STATISTICAL METHODS FOR INTEGRATIVE ANALYSIS OF MULTIPLE TYPES OF DATA

Kin Yau Wong

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2017

Approved by:

Danyu Lin

Donglin Zeng

Michael R. Kosorok

Quefeng Li

Andrew B. Nobel

Charles M. Perou

© 2017
Kin Yau Wong
ALL RIGHTS RESERVED

ABSTRACT

Kin Yau Wong: Statistical Methods for Integrative Analysis of Multiple Types of Data
(Under the direction of Danyu Lin and Donglin Zeng)

Recent technological advances have made it possible to collect multiple types of genomic data on the same set of subjects. It is of great interest to integrate multiple genomic data types to address various imperative biological problems, such as to understand the mechanisms of complex diseases, to predict disease outcomes, and to classify patients into risk groups. Multi-platform genomic data present unprecedented opportunities to address the above issues but also pose significant statistical and computation challenges that cannot be adequately tackled using existing methods.

In the first part of the dissertation, we develop a structural equation modeling (SEM) framework to jointly model the relationships among various types of genomic variables and phenotypes. We extend the current SEM framework to include semiparametric transformation models for potentially right-censored time to event. We develop an efficient nonparametric maximum likelihood estimation approach to estimate the model parameters.

In the second part of the dissertation, we develop a robust score test based on imputed data for testing the association between a phenotype and a partially missing genomic variable. We impute the missing values based on a semiparametric model for the genomic variable against other genomic and non-genomic variables. The proposed test preserves the type I error even when the imputation model is misspecified. We develop a spline-based method to estimate the semiparametric imputation model.

In the third part of the dissertation, we develop a method based on statistical boosting and penalized estimation for the prediction of a potentially right-censored time to event using multiple types of high-dimensional genomic variables. The proposed method properly accounts for the differences in size and predictive power across data types. We propose an efficient and stable computation method that is based on existing algorithms for penalized estimation.

In all three parts of the dissertation, we evaluate the proposed methods using extensive simulation studies and demonstrate their advantages over existing methods. We provide applications for each proposed method to a large-scale multi-platform genomic study, The Cancer Genome Atlas.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude and appreciation to my advisors, Drs. Danyu Lin and Donglin Zeng. It is hard to overstate how much I have learned from them. It has been an amazing yet challenging journey working with them on various research projects, from the initial formulation of problems to the development of methods to the presentation of results. Their intellectual rigor, their knowledge in biostatistics and related fields, their ability to identify important areas of research, and their attentiveness to details inspire me and propel me to strive for improvement and greater achievement.

I would also like to thank Dr. Charles M. Perou for supporting me and offering me the opportunity to work at the Perou Lab since 2015. It is an enriching and humbling experience to work with people from areas of expertise different from mine. I have learned a lot about cancer biology and genomics, as well as conducting scientific research in general.

Moreover, I would like to thank my other committee members, Drs. Michael R. Kosorok, Quefeng Li, and Andrew B. Nobel. They have reviewed this dissertation carefully and provided insightful comments at various stages of the preparation of this dissertation.

Besides, I am grateful that in my first two years of PhD studies, I had the opportunity to work with Dr. Jason P. Fine on an interesting research project. Although the work does not end up in this dissertation, the knowledge and skills I have acquired through the process have been and will continue to be useful for my research.

Furthermore, I would like to thank the Croucher Foundation from Hong Kong for the financial support in my first three years of PhD studies. The generous scholarship took care of all my financial needs and allowed me to work single-mindedly on research.

Last but not least, I would like to thank my parents and all my friends, including colleagues from the Department of Biostatistics and the Perou Lab, as well as those outside my academic circle. They have supported and inspired me in various ways and enriched my otherwise plain and humdrum five years of life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Integrative Analysis of Genomic Data

In the past two decades, there have been revolutionary advances in high-throughput genome-profiling technology. As a result, large-scale genomics studies have been able to collect multiple types of genomic data, including DNA sequences, RNA expressions, and protein expressions, for a large number of subjects. Such multi-platform genomic data provide opportunities for researchers to understand gene functions, identify interactions between and within different types of genomic features, and characterize molecular abnormalities associated with genetic diseases in unprecedented depth and precision. *Integrative* statistical methods have been the essential tools for deciphering these complex multi-platform genomic data.

In this dissertation, we define integrative analysis (of genomic data) broadly as the analysis that combines multiple types of genomic data under a unified framework. This is sometimes referred to as "vertical integration", to distinguish from "horizontal integration" or meta-analysis, which is the combined analysis of subjects from different studies with the same type of data. Distinct from the analysis of a single data type, integrative analysis attempts to understand biological mechanisms as one "integrated system". The main objectives of integrative analysis include: (a) to characterize the complex relationships among different types of molecular structures and their relationships with various phenotypes, (b) to understand the taxonomy of diseases and classify patients into disease subtypes, and (c) to predict clinical outcomes, such as time to disease progression or death, for patients (Kristensen et al. 2014).

Integrative analysis can be more effective in addressing the aforementioned issues than the analysis of a single data type for several reasons. First, any biological mechanism involves the interplay of various layers of genomic features, and genetic diseases usually involve disruptions at multiple molecular levels. The activities of one type of genomic feature only reflect a small aspect of a biological process. For example, the amount and activities of a protein, which play important

roles in biological processes, are not fully determined by the expression of the corresponding RNA nor the coding DNA sequence. Therefore, it takes a combination of data types to characterize a disease and inform accurate phenotype prediction.

Second, the functions of a type of genomic feature can be better understood by evaluating their interactions with other types of genomic features. Some genomic features, such as alterations of DNA, affect the phenotypes of interest indirectly through altering other genomic features. Using multiple genomic data types, we can identify the direct effects of genomic features on a phenotype, as well as their indirect effects through other genomic features. Also, we can identify spurious associations between a genomic feature and a phenotype by controlling for the confounding effects of other genomic features.

Third, information obtained across data types can be aggregated to improve statistical efficiency. Due to technological limitations and the dynamic nature of biological processes, each type of genomic data is subject to noise. The presence of multiple data types allows us to confirm or refine the findings of one data type using other data types. Also, missing data in one data type can be inferred from the observed data using the known biological associations among them. For example, it has been shown that a substantial proportion of variation in gene expression can be explained by point mutations and copy number variations (Stranger et al. 2007).

Despite its advantages, integrative analysis presents unique statistical challenges. First and foremost, the dimensionality and complexity of the problem increase as more data types are included. Each genomic data type is high-dimensional, and the relationships among them are complex. As a result, inclusion of new data types greatly increases the degrees of freedom of the underlying models and may yield weaker inference. New statistical methods that can handle high dimensionality and utilize the known biological relationships among the variables are warranted. Also, multi-platform genomic data typically contain a substantial amount of missing data, because some types of genomic data are too expensive to be measured for a large number of subjects. We need valid and efficient methods to handle missing data.

1.2 Types of Genomic Data

In this section, we describe some commonly studied types of genomic data.

DNA Alteration DNA is the hereditary material in a cell, which contains information for building and maintaining an organism. Genetic diseases, such as cancer, are driven by abnormal changes in the DNA sequences, so that the diseased cells no longer perform their normal functions. DNA alterations that are not inherited from a parent and would not be passed to an offspring are called somatic DNA alterations. DNA alterations can take several forms, including small mutations, such as single nucleotide polymorphisms (SNPs) and short insertion or deletion (indels), and other structural changes, such as copy number variation (CNV).

Methylation Methylation refers to the addition of a methyl group to a DNA molecule. Methylation changes the activities of the DNA without changing the DNA sequence and thus is a form of *epi*genetics. The effect of methylation on gene expression depends on the location; methylations located at gene promoter regions typically suppress gene expression, while methylations located at genic regions typically promote gene expression.

Gene Expression and miRNA Expression A gene is a region of DNA that encodes proteins; there are around 25,000 genes in the human genome. Through the process of transcription, information of the DNA sequence of a gene is used to produce RNA. One type of RNA, called messenger RNA (mRNA), is then further translated into protein; gene expression refers to the expression of mRNA.

A micro RNA (miRNA) is a small non-protein-coding RNA molecule. miRNA serves to regulate gene expression through RNA silencing and post-transcriptional regulation. miRNAs are mostly gene expression inhibitors, although there are cases where the expressions of a miRNA and its target are positively correlated.

RNA expression was traditionally measured using microarray platforms, the first widely used high-throughput technology. In recent years, microarray platforms were superseded by RNA sequencing as the major technology to measure RNA expression. Unlike microarray platforms, RNA sequencing is able to discover novel and rare transcripts and detect small variants.

Protein Proteins perform the basic biological functions of an organism. Changes in protein level and structure have been shown to play essential roles in tumor development and progression, some of which are not reflected by genetic changes. The amount and activity of a protein are controlled

by the expression of the corresponding gene and a set of post-translational modifications (PTM). One of the most-studied PTM is phosphorylation, which plays a crucial role in regulating protein activities and interactions among proteins.

A method to measure the expression of proteins and phospho-proteins is reverse phase protein array (RPPA), a high-throughput antibody-based technique. Because of the inherent difficulty of replicating and measuring proteins, current technologies only allow the measurement of protein expressions at a huge cost. As a result, genomic studies have only measured the expression of a small number of proteins on relatively few subjects.

1.3 The Cancer Genome Atlas

All data analyzed in this dissertation are obtained from The Cancer Genome Atlas (TCGA), a large public cancer genomics database. The TCGA project was undertaken by the National Institutes of Health (NIH), jointly led by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The main goal of the project is to discover and characterize all key genomic alterations in the major types and subtypes of cancer. Since 2005, TCGA has generated detailed molecular data for over 11,000 cancer samples with 33 types of cancer. Subjects were measured for multiple types of genomic data, including somatic mutation, copy number variation, methylation, and expressions of miRNA, mRNA, and proteins, using various methods, including microarrays and next-generation sequencing platforms. Figure 1.1 shows the data types collected by TCGA and their potential relationships.

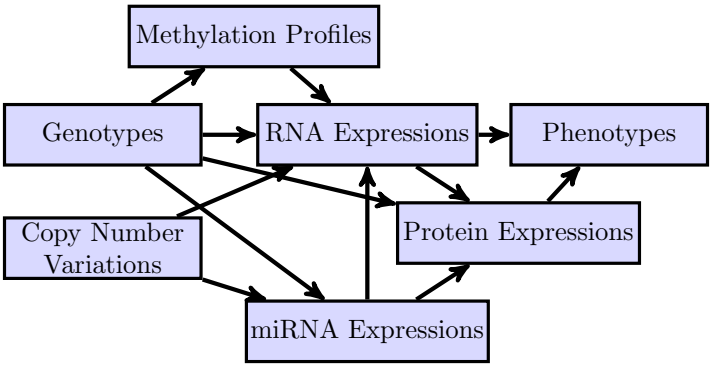


Figure 1.1. Types of Data Measured in TCGA and Their Potential Relationships.

1.4 Existing Statistical Methods for Integrative Analysis

Methods for integrative analysis vary according the type of biological question being addressed. We classify methods for integrative analysis broadly into those that concern (a) the estimation of the relationships among genomic variables and phenotypes, (b) the test of the effects of genomic variables on phenotypes, and (c) the prediction of phenotypes using (high-dimensional) genomic variables.

1.4.1 Modeling Relationships Among Multiple Types of Variables

A popular approach for simultaneously modeling the relationships among multiple variables, both latent and observed, is structural equation modeling (SEM). SEM is traditionally used in the social sciences and psychology research but has also found applications in genomics in recent years. For example, Li et al. (2006) proposed an SEM to capture the association among genetic loci and multiple phenotypes. Lee et al. (2007) proposed an SEM to study the interaction among genes, whose activities were represented by latent variables and are manifested by SNP data. Nock et al. (2007) and Nock et al. (2009) used SEM to study the effects of different genes on rheumatoid arthritis and the Metabolic Syndrome, respectively.

A related approach for jointly modeling different types of genomic variables is PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models) (Vaske et al. 2010). In this approach, quantities that represent observable data, such as copy number variation, gene expression, and protein abundance, and abstract processes, such as DNA damage and apoptosis, are represented by vertices on a graph. The vertices are linked by edges that represent the associations among the quantities. The graph is constructed based on prior biological knowledge. Using an Expectation-Maximization (EM) algorithm, the activity of each unobserved quantity can be inferred. The authors showed that the inferred activity can be used to effectively classify patients into subtypes.

A series of papers by Huang and his collaborators (Huang 2014; 2015; Huang et al. 2014; 2015; 2016) considered the joint analysis of multiple genomic variables and a phenotype, which may be binary, continuous, and censored. Under a mediation analysis framework, the authors were able to identify the direct and indirect effects of genomic variables on other genomic variables and phenotypes.

1.4.2 Testing the Effects of Multiple Types of Variables on Phenotypes

A primary interest in genomics studies is to test the association between genomic variables and phenotypes of interest. For illustration, let $\boldsymbol{G}$ denote a vector of SNP variables from a particular gene or genetic region, $\boldsymbol{S}$ denote a vector of genomic variables of the gene that are regulated by $\boldsymbol{G}$, such as gene expression and protein expression, $\boldsymbol{Z}$ denote a vector of covariates, and Y denote the phenotype of interest. We are interested in testing the association between Y and individual elements of $\boldsymbol{G}$ or the overall activity of the gene. We classify approaches that incorporate information from $\boldsymbol{S}$ into the following categories: (a) methods that combine test results from multiple platforms; (b) methods that incorporate multiple platforms into a single regression framework; (c) methods that separately test the direct and indirect effects of the genetic variants; and (d) methods that incorporate the association between the genetic variants and other platforms.

The first group of approaches combines test results from multiple platforms. For complex phenotypes, the associations between individual genetic variants and the phenotypes are typically weak. If the variables derived from the same genetic region from different platforms exhibit similar association patterns with a phenotype, then one can improve power by combining the test results from the platforms with the results from the genetic variants. Hamid et al. (2009) reviewed methods that combine the effect sizes or p-values from individual platforms. Xiong et al. (2012) developed a method for combining the effects of SNPs and gene expressions, estimated by separate regression analyses, into the effects of the corresponding genes or pathways.

The second group of approaches considers multiple platforms under a single regression framework. Tyekucheva et al. (2011) proposed to fit a standard regression model for a phenotype on genomic variables of a gene from different platforms. A score is given to each gene according to the significance of the joint test of all regression parameters, and the scores from multiple genes are combined using a similar approach as Xiong et al. (2012) into pathway-level association scores. Huang (2014; 2015); Huang et al. (2014; 2015; 2016) considered joint tests of the effects of SNPs and other genomic variables. Instead of adopting a traditional test with the degrees of freedom equal to the number of coefficients being tested, a variance component test was adopted. This approach is powerful if the SNP effects are uniformly spread over the SNPs and the effect of an individual SNP is weak.

The third group of approaches separately tests the direct effect of $\boldsymbol{G}$ on Y or the indirect effect of $\boldsymbol{G}$ on Y through $\boldsymbol{S}$. Typically, the effects of SNPs are mediated through other genomic variables,

such that the direct effects of SNPs are weaker than the indirect effects. As a result, it may be more powerful to isolate and test the indirect effects than testing the marginal effects of the SNPs. Zhao et al. (2014) proposed to test the indirect effect of $\boldsymbol{G}$ by a two-step approach. In the first step, we estimate the function of $\boldsymbol{S}$ through which $\boldsymbol{S}$ affects Y. In the second step, we test the association between $\boldsymbol{G}$ and the function of $\boldsymbol{S}$ identified in the first step; the procedure can be understood as testing the effect of $\boldsymbol{G}$ on the part of Y affected by $\boldsymbol{S}$. This test is powerful when the total effect of $\boldsymbol{G}$ mainly consists of the indirect effect. Huang and Pan (2015) proposed to test the indirect effect of individual components of $\boldsymbol{G}$ through the expression of multiple genes using a variance component test. If the indirect effects through the expression of different genes are of different signs, then the variance component test is more powerful than testing the marginal effects.

Finally, the last group of approaches utilizes the associations among the genomic variables to improve power. One simple method is to screen components of $\boldsymbol{G}$ based on their association with $\boldsymbol{S}$. Because functional SNPs are expected to perturb the expression of genes, SNPs that are not associated with the expression of any gene are likely to be not related to the phenotypes of interest. By screening out the SNPs with very weak or no association with $\boldsymbol{S}$, the burden on multiple-testing correction is lightened, and the overall power is increased. He et al. (2013) developed a Bayesian method, named Sherlock, to search for gene-phenotype associations. The presence of associations between both $\boldsymbol{G}$ and $\boldsymbol{S}$ and $\boldsymbol{G}$ and Y provides positive evidence for gene-phenotype association, while the presence of association between $\boldsymbol{G}$ and $\boldsymbol{S}$ along with the absence of association between $\boldsymbol{G}$ and Y provide negative evidence for gene-phenotype association.

1.4.3 Phenotype Prediction Using Multiple Types of High-Dimensional Variables

A naïve approach to combine information from multiple genomic platforms is to concatenate data matrices from the platforms and perform standard analyses on the combined data. For example, one may perform generic variable selection procedures, such as LASSO (Tibshirani 1996) and elastic net (Zou and Hastie 2005), on the set of predictors from all genomic platforms. However, this approach does not make use of the information that the predictors come from different platforms and are inherently different. We discuss more sophisticated approaches, which can be classified into the following two categories: (a) methods that account for the differences in predictive power of different platforms; and (b) methods that incorporate the associations among variables from different platforms.

The first group of approaches assigns weights to the platforms according to their importance. Lanckriet et al. (2004) adopted a kernel support vector machine (SVM) approach for classification. SVM is a supervised classification procedure that divides the feature space into two compartments, such that observations with different outcomes belong to different compartments and are as separated as possible. The classification rule is computed by solving an optimization problem that involves the phenotype and the distance between each pair of observations in the feature space. The innovation of Lanckriet et al. (2004) is to define the distance between the genomic profiles of two subjects, i.e., the distance in the feature space, by a weighted sum of the distances along the spaces of the genomic platforms. The weights, which are calculated based on the data, reflect the relative importance of the genomic platforms. The special case of zero weight represents that the corresponding platform plays no role in determining the distance between two subjects and thus not affecting the classification rule. Similar approaches were proposed by Daemen et al. (2009) and Seoane et al. (2014) for more complex settings.

The second group of approaches incorporates the associations among genomic platforms. Wang et al. (2013) proposed to decompose gene expression into a component that is regulated by methylation and a component that is not regulated by other variables, where the two components affect the phenotype differently. The Bayesian LASSO was proposed to estimate the effect of each component of the gene expression variables. The advantage of such a decomposition is twofold. First, the gene expression variable can be "denoised", i.e., the effect of the component that is not regulated, which may represent noise, can be set to zero. Second, methylation can be limited to have no direct effect on the phenotype but can have indirect effects through gene expression. This would yield more biologically interpretable results. This approach was extended by Jennings et al. (2013) and Denis and Tadesse (2016) to include more genomic platforms by further decomposition of genomic variables and inclusion of more levels of regulation. Zhu et al. (2016) considered a similar two-step approach. Let $\boldsymbol{S}$ and $\boldsymbol{G}$ be two types of genomic variables, where $\boldsymbol{S}$ is regulated by $\boldsymbol{G}$. First, some low-dimensional linear functions of $\boldsymbol{S}$ that are regulated by $\boldsymbol{G}$ are identified. Then, the phenotype is regressed on the functions of $\boldsymbol{S}$ identified in the first step along with other components of $\boldsymbol{S}$ and $\boldsymbol{G}$. In this approach, overlapped information between $\boldsymbol{G}$ and $\boldsymbol{S}$ can be removed, and $\boldsymbol{S}$ is allowed to have more "direct" effect on the phenotype.

1.5 Outline of Dissertation

In this dissertation, we focus on methods for identifying the associations between genomic variables and phenotypes of interest in integrative genomics studies, with special attention to right-censored time to event. In particular, in the following three chapters, we study the problems of estimation, hypothesis testing, and variable selection in high-dimensional settings, respectively. Future work on variable selection is discussed in the last chapter. The proposed methodology in each chapter is especially designed for integrative analysis, such that they take into account the differences across and relationships among different data types.

In Chapter 2, we consider a semiparametric structural equation modeling framework to incorporate the biological relationships among different types of genomic data and phenotypes, including right-censored time to event. We include latent variables to accommodate measurement error of the genomic variables. We propose a semiparametric approach to efficiently estimate the model parameters.

In Chapter 3, we propose a hypothesis test for the association between a phenotype and a genomic variable with partially missing values. We propose a semiparametric robust approach to impute the missing values using other observed variables. Under general missing-data mechanisms, the type I error of the proposed test is preserved, even when the imputation model is misspecified.

In Chapter 4, we consider a high-dimensional regression model with multiple types of genomic variables as predictors. We propose a method based on statistical boosting and penalized estimation, which accounts for the differences among the different types of variables, for variable selection and phenotype prediction. We develop an algorithm based on the coordinate-descent method for the computation of the solution.

In Chapter 5, we discuss a future direction of research and present some preliminary results. We consider a variable selection problem with partially missing high-dimensional multi-platform genomic data. We propose a joint latent variable model for different types of genomic variables to infer missing values from observed data. We develop a computationally efficient penalization EM algorithm for variable selection.

CHAPTER 2

SEMIPARAMETRIC STRUCTURAL EQUATION MODELING WITH CENSORED DATA

2.1 Introduction

Structural equation modeling (SEM) is a very general and powerful approach to capture complex relationships among multiple factors, both observed and latent (Bollen 1989). A typical SEM framework consists of a structural model that connects latent variables and a measurement model that relates latent variables to observed variables. SEM is extremely popular in the social sciences and psychology, where unmeasured quantities and psychological constructs, such as human intelligence and creativity, can be related to and investigated through observed data. The text of Bollen (1989) has been cited more than 20,000 times. Recently, SEM has gained popularity in medical and public health research (Dahly et al. 2009; Naliboff et al. 2012).

Our interest in SEM was motivated by its potential application to integrative analysis in genomic studies. Recent technological advances have made it possible to collect different types of genomic data, including DNA copy number, SNP genotype, DNA methylation level, and expression levels of mRNA, microRNA, and protein, on a large number of subjects. There is a growing interest in integrating these genomic platforms so as to understand their biological relationships and predict disease progression and death, which are considered potentially censored survival times (The Cancer Genome Atlas (TCGA); https://tcga-data.nci.nih.gov/tcga/).

SEM with discrete survival times has been studied by Rabe-Hesketh et al. (2001; 2004), Muthén and Masyn (2005), and Moustaki and Steele (2005). For continuous survival time, Larsen (2004; 2005) adopted the proportional hazards model (Cox 1972) with a single latent variable to capture the association between the survival time and other observed variables; Asparouhov et al. (2006) considered a more general formulation of the association among the latent and observed variables. SEM with the Cox proportional hazards model for the survival component has been adopted for more complex settings, such as multivariate survival times (Stoolmiller and Snyder 2006) and competing risks (Stoolmiller and Snyder 2013). A popular software program, Mplus (Muthén and Muthén

1998–2015), has implemented SEM with survival data under the proportional hazards model. The estimation of the nonparametric baseline hazard function is based on piecewise-constant splines, and no theoretical justification is available. In fact, the standard error estimator for the baseline hazard function is incorrect.

In this chapter, we propose a general SEM framework that includes a semiparametric component of the measurement model for potentially censored survival times. Specifically, we formulate the effects of latent and observed covariates on survival times through a broad class of semiparametric transformation models that includes the proportional hazards model as a special case. The observed covariates may include manifest variables that depend on latent variables. We study nonparametric maximum likelihood estimation (NPMLE), under which the cumulative hazard functions are estimated by step functions with jumps at observed survival times.

The proposed SEM is reminiscent of joint modeling for survival and longitudinal data (Henderson et al. 2000; Tsiatis and Davidian 2004). With the latter, the observed longitudinal variables are considered error-prone measurements of some underlying latent variables, but the measurements themselves are not causal determinants of the survival time. By contrast, our SEM framework allows latent variables to have direct effects on survival times, as well as indirect effects through other manifest variables. In addition, our framework accommodates much more complex relationships among latent variables.

A major challenge in our theoretical development is model identifiability. Even for an SEM with normally distributed variables, no single set of conditions exists that is both necessary and sufficient for model identifiability. Methods that deal with special cases of the normal SEM were proposed by Bollen (1989), Reilly and O'Brien (1996), Vicard (2000), and Bollen and Davis (2009), among others. Most of the methods are based on the fact that identifiability can be established by solving the equations relating the first two model-implied moments to the sample moments. This approach is not directly applicable to models with nonparametric components, as infinite-dimensional parameters cannot be identified through a finite number of equations. Because the proportional hazards structure results in a likelihood function that takes the form of a Laplace transform, however, we are able to develop sufficient conditions under which the identifiability of a semiparametric SEM can be established by inspecting simpler parametric models.

Another theoretical challenge is the invertibility of the information operator. For the information

operator to be invertible, we require that the score statistic along any non-trivial submodel is non-zero. As in the case of model identifiability, general conditions for the invertibility of the information operator for semiparametric models do not exist. In the existing work involving latent variables for survival times (Kosorok et al. 2004; Zeng and Lin 2010), verifying the invertibility of the information operator involves inspecting the local behavior of the score statistic around the zero survival time. This approach does not make full use of the variability of the score statistic contributed by the survival times and cannot deal with the proposed general modeling framework. We show that the invertibility of the information operator can be verified by inspecting the parametric components of the SEM under some mild conditions in the survival model.

The rest of this chapter is structured as follows. In Chapter 2.2, we formulate the model and describe our approach to establish model identifiability. In Chapter 2.3, we discuss the numerical implementation of the NPMLE. In Chapter 2.4, we present theoretical results for model identifiability and describe the asymptotic properties of the estimators. In Chapter 2.5, we report the results from simulation studies. In Chapter 2.6, we provide an application to the TCGA data, which motivated this work. We make some concluding remarks in Chapter 2.7 and relegate theoretical proofs to Chapter 2.8.

2.2 Basic Framework

2.2.1 Model and Likelihood

Let $\boldsymbol{\eta}$ denote a q-vector of latent variables, $\boldsymbol{Y}$ denote an r-vector of uncensored manifest variables, $(T_1, \ldots, T_K)$ denote K potentially censored survival times, and $\boldsymbol{W}$ and $\boldsymbol{Z}$ denote two vectors of observed covariates. Without loss of generality, assume that the support of the covariates includes zero. We specify the conditional distributions of $\boldsymbol{\eta}$ given $\boldsymbol{Z}$, $\boldsymbol{Y}$ given $\boldsymbol{Z}$ and $\boldsymbol{\eta}$, and T_k given $\boldsymbol{W}$, $\boldsymbol{Z}$, $\boldsymbol{Y}$, and $\boldsymbol{\eta}$ as follows:

$$\boldsymbol{\eta} \mid \boldsymbol{Z} \sim F_\eta\left(\cdot \mid \boldsymbol{Z}; \boldsymbol{\nu}\right), \tag{2.1}$$

$$\boldsymbol{Y} \mid (\boldsymbol{Z}, \boldsymbol{\eta}) \sim F_Y\left(\cdot \mid \boldsymbol{Z}, \boldsymbol{\eta}; \boldsymbol{\psi}\right), \tag{2.2}$$

$$\Lambda_{T_k}\left(t \mid \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\eta}\right) = G_k\{\Lambda_k\left(t\right) e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\}, \; k = 1, \ldots, K, \tag{2.3}$$

where $F_\eta(\cdot \mid \boldsymbol{Z}, \boldsymbol{\nu})$ denotes a q-variate normal distribution function indexed by a parameter vector $\boldsymbol{\nu}$, $F_Y(\cdot \mid \boldsymbol{Z}, \boldsymbol{\eta}; \boldsymbol{\psi})$ denotes an r-variate parametric distribution function indexed by a parameter vector $\boldsymbol{\psi}$, Λ_{T_k} is the cumulative hazard function of T_k given $(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\eta})$, G_k is a known increasing function, Λ_k is an unspecified positive increasing function with $\Lambda_k(0) = 0$, and $(\boldsymbol{\vartheta}_k, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\phi}_k)$ are unknown regression parameters.

Model (2.1) is the structural model of the latent variables. Model (2.2) is the measurement model of $\boldsymbol{Y}$. We assume that $\boldsymbol{Y}$ and $\boldsymbol{\eta}$ are independent of $\boldsymbol{W}$ given $\boldsymbol{Z}$. Models (2.1) and (2.2) represent the existing SEM framework with $\boldsymbol{Y}$ not restricted to be normally distributed. Equation (2.3) includes the proportional hazards and proportional odds models as special cases with the choices of $G_k(x) = x$ and $G_k(x) = \log(1 + x)$, respectively. The proportional hazards model has been considered in the literature.

The survival time T_k is subject to right censoring by C_k. It is assumed that $(C_1, \ldots, C_K)$ are independent of $(T_1, \ldots, T_K)$ and $\boldsymbol{\eta}$ conditional on $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$. Define $\tilde{T}_k = \min(T_k, C_k)$ and $\Delta_k = I(T_k \leq C_k)$, where $I(\cdot)$ is the indicator function. For a sample of size n, the observed data consist of $\mathcal{O}_i \equiv (\tilde{T}_{1i}, \ldots \tilde{T}_{Ki}, \Delta_{1i}, \ldots, \Delta_{Ki}, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)$ $(i = 1, \ldots, n)$.

Let $\boldsymbol{\theta}$ denote the collection of all Euclidean parameters, and write $\mathcal{A} = (\Lambda_1, \ldots, \Lambda_K)$. The likelihood function for $\boldsymbol{\theta}$ and $\mathcal{A}$ is proportional to

$$
\begin{aligned}
L_n(\boldsymbol{\theta}, \mathcal{A}) \;=\; & \prod_{i=1}^{n} \int \prod_{k=1}^{K} \Bigg[\lambda_k(\tilde{T}_{ki}) e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\vartheta}_k + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\phi}_k} \\
& \times G_k' \Big\{ \Lambda_k(\tilde{T}_{ki}) e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\vartheta}_k + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\phi}_k} \Big\} \Bigg]^{\Delta_{ki}} \\
& \times \exp\left[-G_k \Big\{ \Lambda_k(\tilde{T}_{ki}) e^{\boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\vartheta}_k + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\phi}_k} \Big\} \right] \\
& \times f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\eta}; \boldsymbol{\psi}) \, f_\eta(\boldsymbol{\eta} \mid \boldsymbol{Z}_i; \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\eta},
\end{aligned}
\tag{2.4}
$$

where $f'(x) = \mathrm{d}f(x)/\mathrm{d}x$ for any function f, $\lambda_k = \Lambda_k'$, $f_Y = F_Y'$, and $f_\eta = F_\eta'$. The NPMLE is defined to be the maximizer of $L_n(\boldsymbol{\theta}, \mathcal{A})$, in which Λ_k is treated as a step function with jumps at $\tilde{T}_{ki}$ with $\Delta_{ki} = 1$ $(i = 1, \ldots, n)$.

2.2.2 Model Identifiability

We describe our approach to establish model identifiability in this section and defer the technical details to Chapter 2.4. The identifiability results can be summarized by two simple rules. Suppose that we have arranged the survival times such that for some $0 \leq K_1 \leq \min(q, K)$, each of $(T_1, \ldots, T_{K_1})$ regresses on and only on one latent variable and a set of covariates that are independent of the latent variables. (We allow $K_1 = 0$ if no survival time satisfies the given conditions, in which case Rule 1 below is vacuous.) We call an observed variable X an indicator of a latent variable η if X follows a generalized linear model with η as a covariate and is independent of all other manifest variables and survival times conditional on η. We have the following rules:

Rule 1. The latent variables attached to $(T_1, \ldots, T_{K_1})$ can be treated as observed if each of $(T_1, \ldots, T_{K_1})$ depends on at least one observed covariate.

Rule 2. If each latent variable has a separate continuous indicator and the distributions of the latent variables and the indicators are identifiable, then the whole model is identifiable.

To illustrate the usefulness of the two identifiability rules, we present two examples.

Example 2.1. Consider the model depicted in Figure 2.1. In the model, Y_1, Y_2, and Y_3 are conditionally independent normal manifest variables of η, and T is a survival time that follows the proportional hazards model with covariate η. Assume that the regression parameter of Y_1 on η is fixed to be one, $\mathrm{E}(\eta) = 0$, and the regression parameters of η in the models of Y_2 and Y_3 are non-zero.

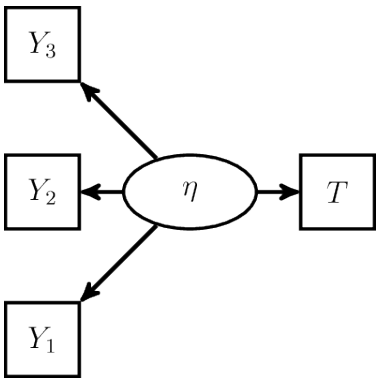


Figure 2.1. The First Example of SEM to Illustrate the Identifiability Rules. The SEM consists of one latent variable, one survival time, and three conditionally independent normal manifest variables.

The above model is similar to the joint model for survival and longitudinal variables. By Bollen (1989)'s three-indicator rule, the model of (Y_1, Y_2, Y_3, η) is identifiable. With Y_1 serving as an indicator of η, Rule 2 implies that the remaining parameters are identifiable. Note that Rule 1 is not applicable in this case because T does not depend on an independent covariate. In fact, the model is not identifiable without (Y_1, Y_2, Y_3) because the scale of the baseline hazard function and the variance of η cannot be separated.

Example 2.2. Consider the model depicted on the left-hand side of Figure 2.2. In the model, Y_1, Y_2, and Y_3 are conditionally independent normal manifest variables of η_2, T_1 is a survival time that follows the proportional hazards model with covariates W and η_1, and T_2 is a survival time that follows the proportional hazards model with covariates η_1 and η_2. Assume that W and Z are non-constant and linearly independent, the regression parameters for the latent variables in the models of T_1 and Y_1 are fixed to be one, $\mathrm{E}(\eta_1) = \mathrm{E}(\eta_2) = 0$, and the regression parameters of W in the model of T_1 and η_2 in the models of Y_2 and Y_3 are non-zero.

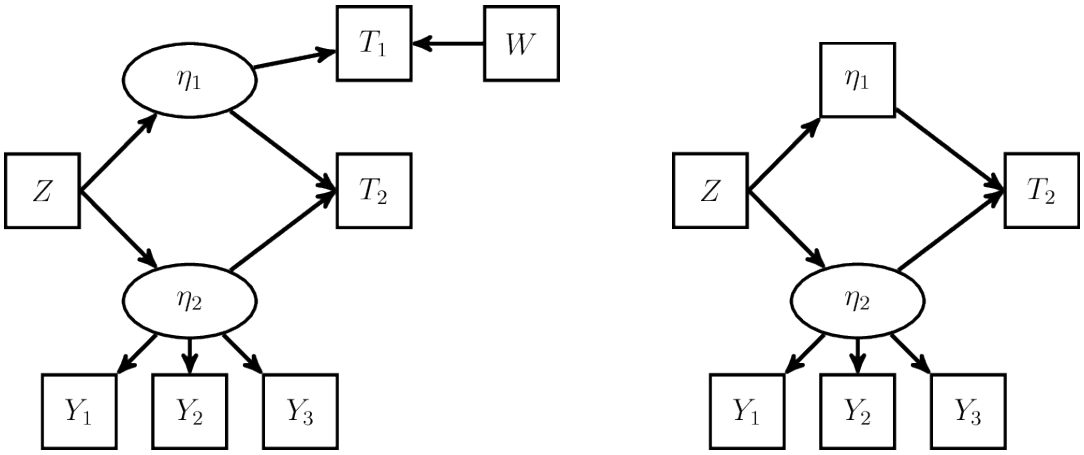


Figure 2.2. The Second Example of SEM to Illustrate the Identifiability Rules. The left panel is an SEM that consists of two latent variables, two observed covariates, two survival times, and three conditionally independent normal manifest variables. The right panel is an intermediate step in identifying the SEM on the left.

First, we use T_1 to help identify the latent variable distributions. By Rule 1, η_1 can be treated as observed when identifying the model. The problem thus reduces to identifying the model shown on the right-hand side of Figure 2.2. The model can then be shown identifiable by the arguments used in Example 2.1.

2.3 Computation of the NPMLE

In this section, we use $\boldsymbol{Z}$ to denote both $\boldsymbol{W}$ and $\boldsymbol{Z}$ with $\boldsymbol{\beta}_k$ $(k = 1, \ldots, K)$ as the corresponding vector of regression parameters. Application of a transformation G_k can be viewed as inclusion of an extra latent variable $\log s_k$ in the regression equation, where s_k is a random variable with density g_k such that $G_k(x) = -\log \int_0^\infty e^{-xt} g_k(t)\, \mathrm{d}t$. We adopt the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) by treating the latent variables, including those introduced by the transformations, as missing data. We perform occasional Newton-Raphson steps to speed up the convergence.

In the combined algorithm, either an EM step or a Newton-Raphson step is performed at each iteration. To avoid confusion, we call the latter an outer Newton-Raphson step. For an EM step, note that the conditional expectation for any function φ of $(\boldsymbol{\eta}_i, \boldsymbol{s}_i) \equiv (\boldsymbol{\eta}_i, s_{1i}, \ldots, s_{Ki})$ given the observed data is

$$\begin{aligned}
\mathrm{E}\left\{\varphi\left(\boldsymbol{\eta}_i, \boldsymbol{s}_i\right) \mid \mathcal{O}_i\right\} &= \mathcal{C}^{-1} \int\int \varphi\left(\boldsymbol{\eta}, \boldsymbol{s}\right) \prod_{k=1}^{K} \Big(\left[\Lambda_k\{\tilde{T}_{ki}\} s_k e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\right]^{\Delta_{ki}} \\
&\quad \times \exp\big(- \int_0^{\tilde{T}_{ki}} s_k e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\, \mathrm{d}\Lambda_k\left(t\right) \big) \Big) \\
&\quad \times f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\eta}; \boldsymbol{\psi}) f_\eta(\boldsymbol{\eta} \mid \boldsymbol{Z}_i; \boldsymbol{\nu}) g\left(\boldsymbol{s}\right)\, \mathrm{d}\boldsymbol{\eta}\, \mathrm{d}s_1 \cdots \mathrm{d}s_K,
\end{aligned}$$

where $\Lambda_k\left\{t\right\}$ is the jump size of the step function Λ_k at t, $\boldsymbol{s} = (s_1, \ldots, s_K)$, $g\left(\boldsymbol{s}\right) = \prod_{k=1}^{K} g_k(s_k)$, and $\mathcal{C}$ equals the above integral evaluated at $\varphi(\cdot, \cdot) = 1$. We use the Gauss-Hermite quadrature to approximate the integrals. To reduce the number of abscissas, we adopt an adaptive quadrature approach (Liu and Pierce 1994). Denote the approximation of the conditional expectation as $\hat{\mathrm{E}}(\cdot)$. After taking expectation on the functions involved, we update $(\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\phi}_k)$ by the one-step Newton-Raphson algorithm on

$$\sum_{i=1}^{n} \log \left[\frac{\hat{\mathrm{E}}\left\{ s_{ki} \exp\left(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}_i^{\mathrm{T}}\boldsymbol{\phi}_k\right)\right\}}{\sum_{j=1}^{n} I(\tilde{T}_{kj} \geq \tilde{T}_{ki}) \hat{\mathrm{E}}\left\{ s_{kj} \exp\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_j^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}_j^{\mathrm{T}}\boldsymbol{\phi}_k\right)\right\}} \right]^{\Delta_{ki}}.$$

Then, we update the cumulative baseline hazard function by

$$\hat{\Lambda}_k\left\{T_{ki}\right\} = \frac{\Delta_{ki}}{\sum_{j=1}^{n} I(\tilde{T}_{kj} \geq \tilde{T}_{ki}) \hat{\mathrm{E}}\left\{ s_{kj} \exp\left(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_j^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}_j^{\mathrm{T}}\boldsymbol{\phi}_k\right)\right\}},$$

where $(\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\phi}_k)$ are evaluated at the current estimates. In addition, we update the remaining parameters at the maximum of

$$\sum_{i=1}^{n} \hat{\mathrm{E}} \left[\log \left\{ f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\eta}_i; \boldsymbol{\psi}) f_\eta(\boldsymbol{\eta}_i \mid \boldsymbol{Z}_i; \boldsymbol{\nu}) \right\} \right].$$

If a closed-form solution is not available, then we apply the one-step Newton-Raphson algorithm to the above expression instead. The above algorithm can be generalized to the case where components of $\boldsymbol{Y}_i$ that do not appear in the model of the survival times are missing at random for some subjects. In this case, we simply drop the corresponding f_Y terms in the evaluation of $\hat{\mathrm{E}}$ and the complete-data log-likelihood.

For an outer Newton-Raphson step, we apply the one-step Newton-Raphson algorithm directly to the logarithm of $L_n(\boldsymbol{\theta}, \mathcal{A})$ given in (2.4) using a similar adaptive quadrature approximation. At the current estimates, the first derivative of $\log L_n(\boldsymbol{\theta}, \mathcal{A})$, i.e., the score statistic, is the same as the first derivative of the expected complete-data log-likelihood. The Hessian matrix used in the Newton-Raphson algorithm can be obtained by Louis (1982)'s formula.

To determine whether an EM step or an outer Newton-Raphson step is to be performed, we keep track of the difference in the log-likelihood at the previous iteration, either an EM or an outer Newton-Raphson step, and the difference at the previous outer Newton-Raphson step. For each iteration, if the log-likelihood difference at the previous step is too small relative to that at the previous outer Newton-Raphson step, then an outer Newton-Raphson step is performed; otherwise, an EM step is performed. Upon convergence, Louis (1982)'s formula is used to obtain the information matrix for the estimation of the standard errors.

The reason that we use a combination of the EM and Newton-Raphson algorithms instead of the Newton-Raphson algorithm alone is two-fold. First, EM steps are more stable, which is important, especially in early iterations. Second, in the estimation of the survival model under the EM algorithm, the regression parameters can be obtained by maximizing the partial-likelihood-type function, and the estimators of the baseline hazard functions take the form of the Breslow estimator. Unlike the Newton-Raphson algorithm, the EM algorithm does not involve the inversion of a high-dimensional matrix.

2.4 Theoretical Properties

2.4.1 Identifiability Conditions

As discussed in Chapter 2.2, we set aside K_1 survival times, $T_1, \ldots, T_{K_1}$, that are used to identify the distribution of the underlying latent variables. We assume that $\text{span}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{K_1}) = \mathbb{R}^{K_1}$. We can choose the K_1 survival times such that each is associated with a few, preferably only one, latent variables. (K_1 is allowed to be 0, in which case we rely solely on the manifest variable $\boldsymbol{Y}$ to identity the latent variable distribution.) Without loss of generality, we assume that $\boldsymbol{\phi}_k = \boldsymbol{e}_k$ ($k = 1, \ldots, K_1$), where $\boldsymbol{e}_k$ is a q-vector with 1 at the kth position and 0 elsewhere. This assumption can be satisfied by applying a linear transformation to the latent variables. Effectively, we fix the scale of the first K_1 latent variables, as is common when establishing model identifiability for SEM. We partition $\boldsymbol{\eta}$ into $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, where $\boldsymbol{\eta}_1 \equiv (\eta_{11}, \ldots, \eta_{1K_1})$ consists of the first K_1 components of $\boldsymbol{\eta}$.

We consider the following identifiability conditions. For the "baseline" hazard functions $(\Lambda_1, \ldots, \Lambda_K)$, we only require identifiability on $[0, \tau]$, where τ denotes the study duration.

(C1) If $(1, \boldsymbol{W}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{c} = 0$ almost surely for some vector $\boldsymbol{c}$ of appropriate dimension, then $\boldsymbol{c} = \boldsymbol{0}$. For $k = 1, \ldots, K$, λ_k is continuous and strictly positive on $[0, \tau]$, and there exists a positive and measurable function g_k such that $\exp\{-G_k(t)\} = \int_0^\infty e^{-ts} g_k(s)\, \mathrm{d}m_k(s)$, where m_k is the Lebesgue measure or the counting measure at 1.

(C2) For $k = 1, \ldots, K_1$, $\mathrm{E}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k \mid \boldsymbol{Z} = \boldsymbol{0}) = 0$, and $\mathrm{E}(\boldsymbol{\eta} \mid \boldsymbol{Z} = \boldsymbol{0}) = \boldsymbol{0}$. Also, for any vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ of appropriate dimensions, $\mathrm{E}(e^{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{c}_1 + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{c}_2} \mid \boldsymbol{Z})$ is finite almost surely.

(C3) For $k = 1, \ldots, K_1$, $\boldsymbol{\vartheta}_k$ is non-zero, and $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are zero.

(C4) Consider two sets of parameters $(\boldsymbol{\psi}, \boldsymbol{\nu})$ and $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$. Let f_{Y,η_1} be the density of $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$ given $\boldsymbol{Z}$. Then, $f_{Y,\eta_1}(\boldsymbol{Y}, \boldsymbol{\eta}_1 \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu}) = f_{Y,\eta_1}(\boldsymbol{Y}, \boldsymbol{\eta}_1 \mid \boldsymbol{Z}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ for all $\boldsymbol{Z}$, $\boldsymbol{Y}$, and $\boldsymbol{\eta}_1$ implies that $\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}$ and $\boldsymbol{\nu} = \tilde{\boldsymbol{\nu}}$.

(C5) Let $(\boldsymbol{Y}^+, \boldsymbol{\eta}_1^+)$ be the components of $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$ that appear in the regression of T_k for some $k = K_1 + 1, \ldots, K$, and let $(\boldsymbol{Y}^-, \boldsymbol{\eta}_1^-)$ be the remaining components. Let $F_{\eta_2|Y,\eta_1}$ be the distribution function of $\boldsymbol{\eta}_2$ given $(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\eta}_1)$ with $(\boldsymbol{Y}^-, \boldsymbol{\eta}_1^-)$ treated as a parameter vector. Then, $\boldsymbol{\eta}_2$ is complete sufficient in $\{F_{\eta_2|Y,\eta_1}(\cdot \mid \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\eta}_1) : \boldsymbol{Z} = \boldsymbol{z}_0, \boldsymbol{Y}^+ = \boldsymbol{y}_0, \boldsymbol{\eta}_1^+ = \boldsymbol{\eta}_{10}\}$ for any fixed $\boldsymbol{z}_0$, $\boldsymbol{y}_0$, and $\boldsymbol{\eta}_{10}$.

Remark 2.1. Condition (C1) pertains to basic requirements on the covariates, the baseline hazard functions, and the transformation functions such that the survival model with observed covariates is identifiable. If m_k is a point mass at 1, then G_k is simply the identity function. Condition (C2) fixes the location parameters of the latent variables and the manifest variables that appear in the regression models of the first K_1 survival times. Condition (C3) requires that the first K_1 survival times depend only on their corresponding latent variable and $\boldsymbol{W}$. The presence of a covariate besides the latent variable is necessary for distinguishing the contributions of the baseline hazard function and the latent variable to the distribution of a survival time that follows a mixture distribution. Condition (C4) requires that the model with observed $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$ is identifiable. Condition (C5) requires that $\boldsymbol{\eta}_2$ is complete sufficient conditional on $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$, where components of $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$ that do not appear in the regression of T_k $(k = K_1 + 1, \ldots, K)$ are treated as parameters, and the rest are held fixed. Conditions (C2) and (C3) are vacuous if $K_1 = 0$, and condition (C5) is vacuous if $K_1 = K$.

We have the following identifiability result.

Theorem 2.1. *Under conditions (C1)-(C5), the model specified by* (2.1)-(2.3) *is identifiable.*

Remark 2.2. The condition that $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are zero for $k = 1, \ldots, K_1$ separates the first K_1 survival times from the remaining observed variables that are associated with the latent variables. This condition is used to simplify the presentation of the identifiability conditions. In the proof of Theorem 2.1, we consider generalized versions of conditions (C3)-(C5), where $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are allowed to be non-zero.

Remark 2.3. Theorem 2.1 implies that the distribution of the latent variable underlying a given survival time can be completely identified if the survival time only regresses on the latent variable and a set of independent covariates. Thus, the survival times make it easy to identify the model, as only a single survival time is enough to identify an underlying latent variable. By contrast, this property does not hold for normal random variables.

Remark 2.4. The derivation of model identifiability from condition (C5) utilizes the property of complete sufficient statistics. The derivation is applicable to general latent variable models; the general result is given by Lemma 2.1 in Chapter 2.8. Lemma 2.1 allows for the establishment of model identifiability by inspecting just a part of the model. It includes Reilly and O'Brien (1996)'s

side-by-side rule, which states that the loadings of an observed variable on any number of latent variables are identifiable if each of the latent variables is attached to a separate independent observed variable whose distribution is identifiable, as a special case.

2.4.2 Asymptotic Properties

Let d be the dimension of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, and Λ_{0k} denote the true value of Λ_k $(k = 1, \ldots, K)$. We impose the following conditions.

(D1) The parameter $\boldsymbol{\theta}_0$ lies in the interior of a compact set $\Theta \subset \mathbb{R}^d$, and the function Λ_{0k} is continuously differentiable with $\lambda_{0k}(t) \equiv \Lambda'_{0k}(t) > 0$ on $[0, \tau]$ for each $k = 1, \ldots, K$.

(D2) With probability one, $P(\tilde{T}_{ki} = \tau \mid \boldsymbol{W}, \boldsymbol{Z}) > \delta_0$ $(k = 1, \ldots, K)$ for some fixed $\delta_0 > 0$.

(D3) Consider any fixed $\boldsymbol{Z}$ and $(\boldsymbol{\psi}, \boldsymbol{\nu}) \in \Theta_{\psi\nu}$, where $\Theta_{\psi\nu}$ consists of the $(\boldsymbol{\psi}, \boldsymbol{\nu})$-component of every $\boldsymbol{\theta} \in \Theta$. For any constant $a_1 > 0$ and $\delta = 0, 1$,

$$\mathrm{E}\left\{\int e^{a_1(1+|\boldsymbol{Y}|+|\boldsymbol{\eta}|)} f_Y\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\eta}; \boldsymbol{\psi}\right)^\delta f_\eta\left(\boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\nu}\right) \mathrm{d}\boldsymbol{\eta}\right\} < \infty.$$

Also, for $j = 1, 2, 3$, there exists a constant $a_2 > 0$ such that

$$\left|\frac{\frac{\partial^j}{\partial \boldsymbol{\psi}^j} f_Y\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\eta}; \boldsymbol{\psi}\right)}{f_Y\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\eta}; \boldsymbol{\psi}\right)}\right| + \left|\frac{\frac{\partial^j}{\partial \boldsymbol{\nu}^j} f_\eta\left(\boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\nu}\right)}{f_\eta\left(\boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\nu}\right)}\right| \leq e^{a_2(1+|\boldsymbol{Y}|+|\boldsymbol{\eta}|)}.$$

In addition, for some positive constants M_j and c_j, $\boldsymbol{N}_j \in \mathbb{R}^r$, and $\varphi_1 \in \ell^\infty(\mathbb{R}^r)$,

$$e^{\sum_{j=1}^q -M_j|b_j|} \leq e^{\sum_{j=1}^q (\boldsymbol{N}_j^{\mathrm{T}} \boldsymbol{Y} b_j + c_j b_j^2)} \varphi_1\left(\boldsymbol{Y}\right) f_Y\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\eta}; \boldsymbol{\psi}\right) f_\eta\left(\boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\nu}\right) \leq e^{\sum_{j=1}^q M_j|b_j|},$$

where $\boldsymbol{b} = S\left(\boldsymbol{\eta}\right)$ for some one-to-one linear transformation S.

(D4) The function G_k is four-times differentiable, $G_k\left(0\right) = 0$, and $K_{1k}\left(1+x\right)^{-\kappa_{1k}} \leq G'_k\left(x\right) \leq K_{2k}\left(1+x\right)^{\kappa_{2k}}$ for some positive constants κ_{1k}, κ_{2k}, K_{1k}, and K_{2k}. Also, $G_k(x)/x^{\rho_k} \to M_k$ or $G_k(x)/\log(x) \to M_k$ as $x \to \infty$ for some positive constants M_k and ρ_k. In addition, $\exp\left\{-G_k(x)\right\} \leq \mu_k\left(1+x\right)^{-\kappa_{3k}}$ for some μ_k and $\kappa_{3k} > \kappa_{2k} + 1$. Furthermore, for some r_k,

$$\sup_{x \geq 0} \frac{\left|G''_k\left(x\right)\right| + \left|G^{(3)}_k\left(x\right)\right| + \left|G^{(4)}_k\left(x\right)\right|}{G'_k(x)\left(1+x\right)^{r_k}} < \infty,$$

where G_k'' and $G_k^{(j)}$ denote the second and jth derivatives of G_k, respectively.

(D5) Let $(\boldsymbol{Z}^{(k)}, \boldsymbol{Y}^{(k)})$ be the components of $(\boldsymbol{Z}, \boldsymbol{Y})$ that appear in the regression of T_k $(k = 1, \ldots, K_1)$. For any vectors $\boldsymbol{h}_1$, $\boldsymbol{h}_{2k}$, and $\boldsymbol{h}_{3k}$ of appropriate dimensions, if

$$\frac{\partial}{\partial(\boldsymbol{\psi}, \boldsymbol{\nu})} f_{Y,\eta_1}(\boldsymbol{Y}, \boldsymbol{\eta}_1 \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu})^{\mathrm{T}} \boldsymbol{h}_1 - \sum_{k=1}^{K_1} (\boldsymbol{Z}^{(k)\mathrm{T}} \boldsymbol{h}_{2k} + \boldsymbol{Y}^{(k)\mathrm{T}} \boldsymbol{h}_{3k}) \frac{\partial}{\partial \eta_{1k}} f_{Y,\eta_1}(\boldsymbol{Y}, \boldsymbol{\eta}_1 \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu})$$

is equal to 0 for all $\boldsymbol{Z}$, $\boldsymbol{Y}$, and $\boldsymbol{\eta}_1$, then $\boldsymbol{h}_1 = \boldsymbol{0}$, $\boldsymbol{h}_{2k} = \boldsymbol{0}$, and $\boldsymbol{h}_{3k} = \boldsymbol{0}$ for $k = 1, \ldots, K_1$, where f_{Y,η_1} is defined in condition (C4).

Remark 2.5. Conditions (D1)-(D4) are similar to the conditions of Zeng and Lin (2010) for joint modeling of longitudinal and survival data. Extra conditions are imposed on the transformations and the distributions of $\boldsymbol{Y}$ and $\boldsymbol{\eta}$ to accommodate the presence of unbounded covariate $\boldsymbol{Y}$ in the survival model. Condition (D5) is for the invertibility of the information operator. If $\boldsymbol{\alpha}_k = \boldsymbol{0}$ and $\boldsymbol{\beta}_k = \boldsymbol{0}$ for $k = 1, \ldots, K_1$, then condition (D5) simply requires that the information matrix of the model for $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$ is invertible. This is parallel to condition (C4) for identifiability.

Remark 2.6. The conditions for identifiability and the invertibility of the information operator (C1)-(C5), and (D5) differ significantly from the corresponding conditions (C5) and (C7) of Zeng and Lin (2010). The latter are stated under very general settings, but they are hard to verify for specific models, especially under our SEM framework. By contrast, our conditions are easier to verify and have intuitive interpretations. For the model in Example 2.1, (D5) simply requires that the model of (Y_1, Y_2, Y_3, η_1) given Z has a non-zero score statistic, which clearly holds.

Let $\mathcal{A}_0 = (\Lambda_{01}, \ldots, \Lambda_{0K})$ and $(\hat{\boldsymbol{\theta}}, \hat{\mathcal{A}})$ be the NPMLE of $(\boldsymbol{\theta}, \mathcal{A})$. Also, let $\mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^d, |\boldsymbol{v}| \leq 1\}$ and $\mathcal{Q} = \{h(t) : \|h(t)\|_{V[0,\tau]} \leq 1\}$ with $\|\cdot\|_{V[0,\tau]}$ being the total variation norm on $[0, \tau]$. We consider $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \hat{\mathcal{A}} - \mathcal{A}_0)$ as a random element in $l^\infty(\mathcal{V} \times \mathcal{Q}^K)$ with

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \hat{\mathcal{A}} - \mathcal{A}_0)(\boldsymbol{v}, h_1, \ldots, h_K) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \boldsymbol{v} + \sum_{k=1}^{K} \int_0^\tau h_k(s)\, \mathrm{d}(\hat{\Lambda}_k - \Lambda_{0k})(s).$$

We have the following results.

Theorem 2.2. *Under conditions (C1)-(C5) and (D1)-(D5),*

1. $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sum_{k=1}^{K} \sup_{t \in [0,\tau]} |\hat{\Lambda}_k(t) - \Lambda_{0k}(t)| \to_{a.s.} 0$*; and*

2. $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \hat{\mathcal{A}} - \mathcal{A}_0) \to_d \mathcal{G}$ *in* $l^\infty(\mathcal{V} \times \mathcal{Q}^K)$*, where* $\mathcal{G}$ *is a continuous zero-mean Gaussian process. Furthermore, the limiting covariance matrix of* $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *attains the semiparametric efficiency bound.*

Remark 2.7. The proof of Theorem 2.2 relies on the Donsker properties of certain classes of functions. It is more challenging to establish the Donsker results in our setting than in previous settings (e.g., Kosorok et al. (2004) and Zeng and Lin (2010)) because the likelihood function of the proposed model may contain the unbounded variable $\boldsymbol{Y}$.

Remark 2.8. A key step in proving the asymptotic normality of the NPMLE is to show that the information operator is invertible. The result is given by Lemma 2.2 in Chapter 2.8, which states that condition (D5), together with conditions (C1)-(C3), and (C5), implies that the information operator of the model is invertible. With this result, we can verify the invertibility of the information operator of the semiparametric model by inspecting the parametric part of the model that contains the observed and latent variables. For the frailty models in Kosorok et al. (2004), verification of the invertibility of the information operator involves inspection of the local behavior of the score around $T = 0$. However, that approach is limited to frailty distributions that are indexed by a one-dimensional parameter and is not directly applicable to cases with more complex latent variable distributions such as those in our setting.

2.5 Simulation Studies

We considered a model with covariates $\boldsymbol{Z} = (Z_1, Z_2)^{\mathrm{T}}$, two latent variables (η_1, η_2), observed continuous variables $(Y_1, \ldots, Y_5)$, observed binary variables (Y_6, Y_7), and a survival time T. Their distributions are given by

$$
\begin{aligned}
\Lambda_T(t \mid \boldsymbol{Z}, Y_6, Y_7, \eta_2) &= G\{\Lambda_0(t)\exp(\boldsymbol{X}_T^{\mathrm{T}}\boldsymbol{\beta}_T + \phi_T\eta_2)\}, \quad \boldsymbol{X}_T = (Z_1, Z_2, Y_6, Y_7)^{\mathrm{T}}, \\
\text{logit}\,\{\mathrm{P}(Y_6 = 1 \mid \boldsymbol{Z}, \eta_2)\} &= \boldsymbol{X}_{Y_6}^{\mathrm{T}}\boldsymbol{\beta}_{Y_6} + \phi_{Y_6}\eta_2, \quad \boldsymbol{X}_{Y_6} = (1, Z_1, Z_2)^{\mathrm{T}}, \\
\text{logit}\,\{\mathrm{P}(Y_7 = 1 \mid \boldsymbol{Z}, Y_6, \eta_2)\} &= \boldsymbol{X}_{Y_7}^{\mathrm{T}}\boldsymbol{\beta}_{Y_7} + \phi_{Y_7}\eta_2, \quad \boldsymbol{X}_{Y_7} = (1, Z_1, Z_2, Y_6)^{\mathrm{T}}, \\
Y_j \mid \eta_1 &\sim \mathrm{N}(\beta_{Y_j} + \phi_{Y_j}\eta_1, \sigma^2_{Y_j}),\ j = 1, 2, 3, \\
Y_j \mid \eta_2 &\sim \mathrm{N}(\beta_{Y_j} + \phi_{Y_j}\eta_2, \sigma^2_{Y_j}),\ j = 4, 5, \\
\eta_2 &\sim \mathrm{N}(\beta_\eta\eta_1, \sigma^2_{\eta_2}), \\
\eta_1 &\sim \mathrm{N}(0, \sigma^2_{\eta_1}).
\end{aligned}
$$

The parameters ϕ_{Y_1} and ϕ_{Y_4} are fixed to be one. The model is depicted in Figure 2.3.

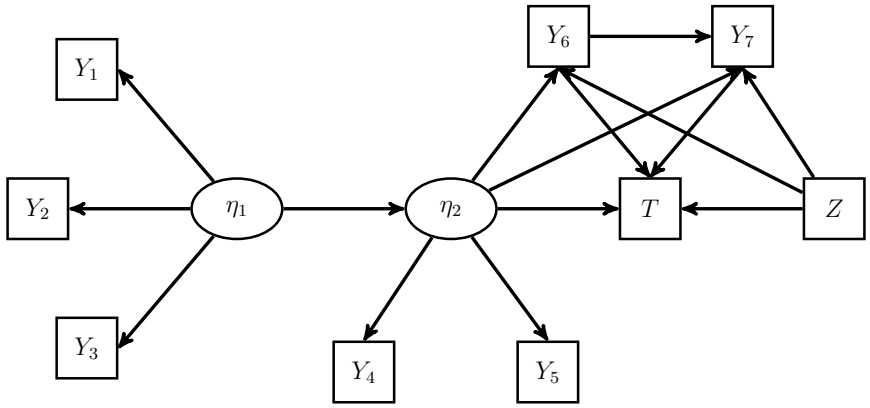


Figure 2.3. Model Used in Simulation Studies. The SEM consists of two latent variables, an observed covariate, seven binary or normal manifest variables, and a survival time that regresses on the latent variable, some manifest variables, and the observed covariates.

We set Z_1 and Z_2 to independent standard normal and Bernoulli(0.5), respectively, and $\Lambda_0(t) = t^2$. We considered the class of logarithmic transformations $G(x) = r^{-1}\log(1 + rx)$ with $r = 0$ or 1, which correspond to the proportional hazards and proportional odds models, respectively. We generated the censoring times from Exp(c), where c was chosen to yield approximately 30% censored observations. We set $(Y_1, \ldots, Y_5)$ to be missing completely at random for 30% of the subjects. We set the sample size to 400 and set the number of abscissa points to 20 for each Gauss-Hermite quadrature. We simulated 5,000 datasets for each setting. The results are summarized in Table 2.1.

The estimators of all parameters are virtually unbiased for both the proportional hazards and proportional odds models. The standard error estimators accurately reflect the true variations, and the coverage probabilities of the confidence intervals are close to the nominal level. Standard error estimators for the parameters in the survival model are larger under the proportional odds model than under the proportional hazards model. As a result, the standard error estimators for the parameters associated with η_2 are larger. The standard error estimators for the remaining parameters are very similar between the two models.

We also evaluated Mplus (Muthén and Muthén 1998–2015) under the proportional hazards model, and the results are presented in Table 2.2. The results for the Euclidean parameters are similar to those presented in Table 2.1. Mplus provides estimator for the baseline hazard function instead of the cumulative baseline hazard function. Its standard error estimator does not reflect the true variation, and the coverages of the confidence intervals are far below the nominal level.

Table 2.1. Simulation Results for the SEM with Two Latent Variables.

		Proportional Hazards Model					Proportional Odds Model				
Dep	Ind	Param	Bias	SE	SEE	CP	Param	Bias	SE	SEE	CP
T	Z_1	0.100	0.003	0.075	0.073	0.947	0.100	0.003	0.109	0.107	0.946
	Z_2	−0.200	−0.005	0.180	0.181	0.953	−0.200	0.002	0.269	0.270	0.950
	Y_6	0.100	0.000	0.145	0.142	0.947	0.100	0.001	0.212	0.208	0.944
	Y_7	0.200	−0.001	0.165	0.163	0.949	0.200	0.002	0.238	0.236	0.950
	η_2	0.500	0.012	0.167	0.166	0.950	0.500	0.014	0.228	0.223	0.948
	$\Lambda_0(t_1)$	0.202	0.004	0.050	0.049	0.949	0.212	0.006	0.074	0.072	0.947
	$\Lambda_0(t_2)$	0.518	0.010	0.120	0.118	0.949	0.608	0.020	0.206	0.202	0.953
	$\Lambda_0(t_3)$	1.103	0.032	0.256	0.249	0.950	1.588	0.076	0.554	0.538	0.950
Y_1	Int	0.000	−0.001	0.073	0.073	0.947	0.000	−0.001	0.073	0.073	0.948
	Var	1.000	−0.010	0.131	0.128	0.960	1.000	−0.010	0.131	0.128	0.960
Y_2	Int	0.000	0.001	0.074	0.073	0.945	0.000	0.001	0.074	0.073	0.945
	η_1	1.000	0.021	0.210	0.203	0.949	1.000	0.022	0.210	0.204	0.950
	Var	1.000	−0.013	0.129	0.128	0.960	1.000	−0.013	0.129	0.128	0.960
Y_3	Int	0.000	0.000	0.073	0.073	0.954	0.000	0.000	0.073	0.073	0.954
	η_1	1.000	0.024	0.211	0.204	0.950	1.000	0.024	0.212	0.204	0.952
	Var	1.000	−0.015	0.131	0.128	0.959	1.000	−0.015	0.131	0.128	0.960
Y_4	Int	0.000	0.001	0.077	0.076	0.946	0.000	0.001	0.077	0.076	0.946
	Var	1.000	−0.025	0.194	0.185	0.956	1.000	−0.029	0.216	0.209	0.951
Y_5	Int	0.000	0.001	0.077	0.076	0.950	0.000	0.001	0.077	0.076	0.949
	η_2	1.000	0.048	0.303	0.279	0.942	1.000	0.060	0.348	0.323	0.945
	Var	1.000	−0.030	0.198	0.188	0.957	1.000	−0.036	0.221	0.213	0.954
Y_6	Int	0.000	0.004	0.278	0.273	0.953	0.000	0.004	0.277	0.273	0.954
	Z_1	−0.500	−0.006	0.113	0.113	0.953	−0.500	−0.006	0.113	0.113	0.953
	Z_2	0.500	0.006	0.301	0.297	0.954	0.500	0.006	0.300	0.297	0.955
	η_2	0.000	0.003	0.229	0.220	0.964	0.000	0.003	0.231	0.222	0.965
Y_7	Int	1.000	0.033	0.347	0.345	0.953	1.000	0.033	0.347	0.345	0.953
	Z_1	1.000	0.024	0.151	0.148	0.951	1.000	0.024	0.151	0.148	0.951
	Z_2	0.200	−0.005	0.344	0.342	0.955	0.200	−0.005	0.345	0.342	0.955
	Y_6	−0.200	−0.009	0.265	0.264	0.952	−0.200	−0.009	0.265	0.264	0.953
	η_2	0.000	0.003	0.265	0.253	0.960	0.000	0.002	0.268	0.255	0.962
η_1	Var	0.500	0.006	0.136	0.134	0.956	0.500	0.006	0.136	0.134	0.955
η_2	η_1	0.500	0.009	0.153	0.147	0.939	0.500	0.009	0.157	0.151	0.940
	Var	0.500	0.009	0.177	0.169	0.960	0.500	0.013	0.196	0.190	0.964

NOTE: Each row corresponds to the regression parameter of the dependent variable "Dep" on the independent variable "Ind" or some other parameter in the model of "Dep". "Int" and "Var" stand for the intercept and error variance, respectively. The parameters $\Lambda_0(t_1)$, $\Lambda_0(t_2)$, and $\Lambda_0(t_3)$ correspond to the cumulative baseline hazard function values at the 25%, 50%, and 75% quantiles of the survival time. The true value of a parameter is given under "Param". "Bias" is the empirical bias; "SE" is the empirical standard error; "SEE" is the empirical mean of the standard error estimator; and "CP" is the empirical coverage probability of the 95% confidence interval.

Table 2.2. Simulation Results for Mplus

Dep	Ind	Param	Bias	SE	SEE	CP
T	Z_1	0.100	0.003	0.075	0.073	0.947
	Z_2	−0.200	−0.005	0.180	0.181	0.953
	Y_6	0.100	0.000	0.145	0.142	0.948
	Y_7	0.200	0.000	0.165	0.163	0.949
	η_2	0.500	0.026	0.172	0.169	0.958
	$\lambda_0(t_1)$	0.900	0.019	2.353	0.939	0.643
	$\lambda_0(t_2)$	1.440	0.016	2.947	1.487	0.689
	$\lambda_0(t_3)$	2.100	0.056	4.615	2.203	0.715
Y_1	Int	0.000	−0.001	0.073	0.073	0.948
	Var	1.000	−0.010	0.131	0.128	0.959
Y_2	Int	0.000	0.001	0.074	0.073	0.944
	η_1	1.000	0.021	0.210	0.203	0.949
	Var	1.000	−0.013	0.129	0.128	0.961
Y_3	Int	0.000	0.000	0.073	0.073	0.954
	η_1	1.000	0.023	0.211	0.203	0.950
	Var	1.000	−0.015	0.131	0.128	0.960
Y_4	Int	0.000	0.001	0.077	0.075	0.945
	Var	1.000	0.001	0.189	0.180	0.942
Y_5	Int	0.000	0.001	0.077	0.076	0.950
	η_2	1.000	0.103	0.346	0.306	0.952
	Var	1.000	−0.056	0.215	0.200	0.962
Y_6	Int	0.000	−0.004	0.278	0.273	0.954
	Z_1	−0.500	−0.006	0.113	0.113	0.953
	Z_2	0.500	0.005	0.301	0.297	0.954
	η_2	0.000	0.003	0.237	0.226	0.961
Y_7	Int	1.000	−0.033	0.347	0.345	0.953
	Z_1	1.000	0.024	0.151	0.148	0.951
	Z_2	0.200	−0.005	0.344	0.342	0.955
	Y_6	−0.200	−0.009	0.265	0.264	0.952
	η_2	0.000	0.003	0.273	0.260	0.960
η_1	Var	0.500	0.006	0.136	0.134	0.955
η_2	η_1	0.500	−0.003	0.155	0.147	0.928
	Var	0.500	−0.018	0.169	0.160	0.953

NOTE: Each row corresponds to the regression parameter of the dependent variable "Dep" on the independent variable "Ind" or some other parameter in the model of "Dep". "Int" and "Var" stand for the intercept and error variance, respectively. The parameters $\lambda_0(t_1)$, $\lambda_0(t_2)$, and $\lambda_0(t_3)$ correspond to the baseline hazard function values at the 25%, 50%, and 75% quantiles of the survival time. The true value of a parameter is given under "Param". "Bias" is the empirical bias; "SE" is the empirical standard error; "SEE" is the empirical mean of the standard error estimator; and "CP" is the empirical coverage probability of the 95% confidence interval.

2.6 Real Data Analysis

We analyzed a dataset on patients with serous ovarian cancer from the TCGA project (The Cancer Genome Atlas Research Network 2011). Genomic variables include DNA copy number, SNP genotype, DNA methylation level, and levels of expression of mRNA, microRNA, total protein, and phosphorylated protein. Demographic and clinical variables include age at diagnosis, race, tumor stage, tumor grade, time to tumor progression, and time to death. There are a total of 586 patients. The median follow-up time was about 2.5 years, and roughly 30% of the patients were lost to follow-up before tumor progression or death. The data are available from http://gdac.broadinstitute.org/.

We focused on the integrative analysis of clinical outcomes and expression levels of mRNA, total protein, and phosphorylated protein. We considered mRNA expression as a latent variable that can only be observed with error through three microarray platforms, namely Agilent 244K Whole Genome Expression Array, Affymetrix HT-HG-U133A, and Affymetrix Exon 1.0. We assumed that the effects of a gene on clinical outcomes are mediated through unobserved protein activity. The latent protein activity is modified by mRNA expression and is manifest through the observed protein expression measurements, which were obtained from the reverse-phase protein arrays platform. Figure 2.4 depicts the SEM fit for each gene. We assumed that the observed variables follow the distributions described in Chapter 2.5, with (Y_1, Y_2, Y_3) being the three microarray measurements, (Y_4, Y_5) = (Total protein expression, Phosphorylated protein expression), (Y_6, Y_7) = (Tumor stage, Tumor grade), (Z_1, Z_2) = (Age, Race), and T being progression-free survival time.

We dichotomized tumor stage into stage II/III versus stage IV and tumor grade into grade 2 versus grade 3/4. Race was dichotomized into white and non-white. We allowed mRNA expression and protein expression data to be missing for some subjects. We excluded patients with tumor stage I or grade 1, as those patients may have a disease that is biologically different from that of patients with tumors of other stages or grades. For each gene, we fit the class of transformation models with $G(x) = r^{-1}\log(1 + rx)$ over a grid of $r = (0, 0.1, \ldots, 2)$. We selected the model with the smallest AIC or, equivalently, the largest log-likelihood value.

We present the results for the gene ACACA. The sample size is 542. About 30% of the subjects do not have protein expression data, and over 10% of the subjects miss at least one mRNA expression measurement. The best-fitting model is obtained at $r = 1$, which corresponds to the proportional

odds model. The point estimates and standard error estimates of the parameters associated with the latent variables are shown in Figure 2.4. The remaining results are shown in Table 2.3. The latent variables have strong positive association with the measurement platforms. As expected, latent protein activity and latent mRNA expression are highly correlated. Latent protein activity is positively associated with progression-free survival time, with a p-value of 0.100. Specifically, higher latent protein activity is associated with shorter progression-free survival time, which agrees with the findings of the literature (Menendez and Lupu 2007). The association of ACACA with tumor stage or tumor grade is weak.

The results for the parameters in the non-survival models are similar between $r = 0$ and 1. The parameters in the survival model have different interpretations between $r = 0$ and 1. With $r = 0$, a unit increase in latent protein activity would have a multiplicative effect of exp(0.068) on the hazard function. With $r = 1$, a unit increase in the latent protein activity would have a multiplicative effect of exp(-0.192) on the survival odds. For this dataset, the proportional odds model provides much stronger evidence for the effect of protein activity on progression-free survival than the proportional hazards model.

For the Cox proportional hazards model, we also present the results from Mplus in Table 2.3. The results from NPMLE and Mplus are similar for most parameters. There are considerable differences between the cumulative baseline hazard function estimates. The standard error estimates for the cumulative baseline hazard function are not available from Mplus.

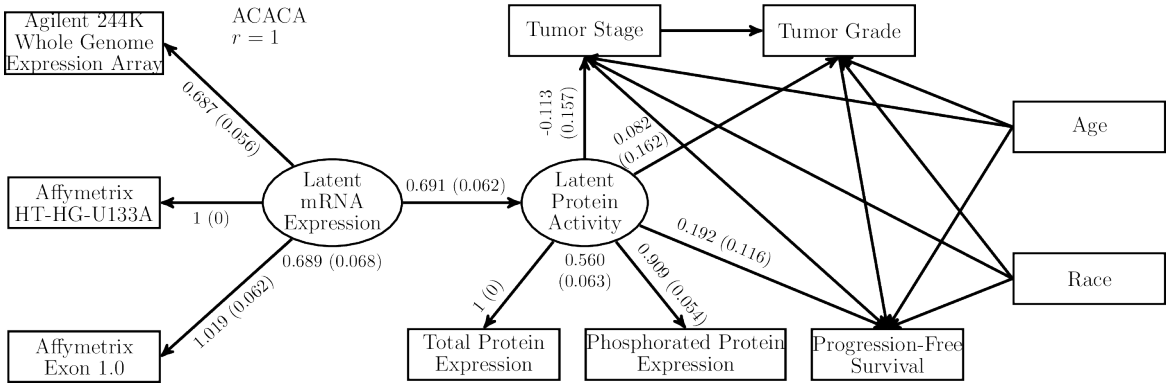


Figure 2.4. Results from the SEM Analysis of the Gene ACACA. Analysis results are from 542 patients with ovarian cancer in the TCGA project. The numbers besides an arrow correspond to the point estimate and standard error estimate (in parentheses) of the regression parameter. The numbers below the latent variables correspond to the point estimate and standard error estimate (in parentheses) of the error variance.

Table 2.3. Analysis Results for the Gene ACACA

		NPMLE ($r = 1$)		NPMLE ($r = 0$)		Mplus ($r = 0$)	
Dep	Ind	Est	St Error	Est	St Error	Est	St Error
T	Z_1	0.229	0.086	0.114	0.054	0.114	0.054
	Z_2	0.038	0.309	0.168	0.192	0.168	0.192
	Y_6	0.760	0.284	0.436	0.139	0.436	0.139
	Y_7	0.511	1.118	0.263	0.151	0.263	0.148
	η_2	0.192	0.116	0.068	0.062	0.068	0.062
	$\Lambda_0(t_1)$	0.135	0.143	0.160	0.028	0.160	N/A
	$\Lambda_0(t_2)$	0.401	0.415	0.392	0.061	0.394	N/A
	$\Lambda_0(t_3)$	1.062	1.061	0.767	0.114	0.771	N/A
Y_1	Int	0.004	0.044	0.004	0.044	0.005	0.044
	Var	0.286	0.038	0.288	0.038	0.287	0.038
Y_2	Int	−0.013	0.044	−0.013	0.044	−0.013	0.044
	η_1	0.687	0.056	0.688	0.056	0.687	0.056
	Var	0.658	0.046	0.658	0.046	0.658	0.046
Y_3	Int	0.014	0.043	0.014	0.044	0.014	0.043
	η_1	1.019	0.062	1.022	0.062	1.021	0.062
	Var	0.276	0.039	0.274	0.039	0.275	0.039
Y_4	Int	−0.021	0.048	−0.023	0.049	−0.022	0.048
	Var	0.096	0.046	0.097	0.046	0.106	0.045
Y_5	Int	−0.009	0.048	−0.011	0.049	−0.011	0.048
	η_2	0.909	0.054	0.910	0.054	0.919	0.054
	Var	0.237	0.041	0.236	0.041	0.229	0.041
Y_6	Int	−1.733	0.128	−1.734	0.128	−1.732	0.128
	Z_1	−0.290	0.125	−0.290	0.125	−0.289	0.125
	Z_2	−0.095	0.434	−0.092	0.434	−0.094	0.434
	η_2	−0.113	0.157	−0.121	0.157	−0.123	0.158
Y_7	Int	1.977	0.149	1.977	0.149	1.996	0.151
	Z_1	0.132	0.135	0.132	0.135	0.133	0.135
	Z_2	0.001	0.461	0.002	0.461	−0.015	0.462
	Y_6	−0.055	0.356	−0.054	0.356	−0.072	0.357
	η_2	0.082	0.162	0.080	0.161	0.081	0.162
η_1	Var	0.689	0.068	0.687	0.068	0.687	0.068
η_2	η_1	0.691	0.062	0.692	0.063	0.690	0.063
	Var	0.560	0.063	0.559	0.063	0.551	0.062

NOTE: Each row corresponds to the regression parameter of the dependent variable "Dep" on the independent variable "Ind" or some other parameter in the model of "Dep". "Int" and "Var" stand for the intercept and error variance, respectively. The representation of each variable is given in Chapter 2.6. The parameters $\Lambda_0(t_1)$, $\Lambda_0(t_2)$, and $\Lambda_0(t_3)$ correspond to the cumulative baseline hazard function values at the 25%, 50%, and 75% quantiles of the progression-free survival, respectively. The point estimate of and standard error estimate of a parameter are given under "Est" and "St Error", respectively.

For comparisons, we also fit a proportional odds model without latent variables for progression-free survival on the covariates and the two protein expression variables, where the subjects with missing protein expression data were discarded. The p-value of the Wald test for the joint effect of protein expression is 0.157. With $r = 0$, the Wald test p-value is 0.578. Therefore, analyses based on standard models fail to conclude a strong association between the protein expression and progression-free survival. The power of the proposed SEM framework stems from the appropriate handling of missing data, the dimension reduction of the observed covariates, and the flexibility of the survival model.

2.7 Discussion

In this chapter, we consider semiparametric SEM for potentially right-censored survival time data. We prove the consistency, asymptotic normality, and semiparametric efficiency of the NPMLE. We propose new rules for establishing model identifiability and invertibility of the information operator. We construct an EM algorithm to compute the NPMLE and introduce occasional Newton-Raphson steps to accelerate the convergence.

One contribution of Theorem 2.1 is that it reduces a semiparametric identifiability problem to a parametric one; it shows that the inclusion of the semiparametric component does not make the model less identifiable but, in some sense, makes the model more easily identifiable. With that being said, the result hinges on correct specification of the model and does not guarantee empirical identifiability in a finite sample. Therefore, care should be taken when fitting a model that is nearly non-identifiable. Another main result of ours is given by Lemma 2.1. This lemma is applicable to a wide range of latent variable models and allows one to deduce the identifiability of a model by inspecting just part of it.

Invertibility of the information operator has received much less attention in the literature than model identifiability. In this chapter, we prove a general result for invertibility of the information operator. It is evident from the proof that the invertibility of the information operator can be established using techniques similar to those used to establish model identifiability. Specifically, the key to the proof of the identifiability of the mixture Cox model is that with the presence of a covariate that is independent of the latent variable, the contributions to the likelihood from the latent variable and the baseline hazard function can be separated by considering different values of the covariate. (In a normal mixture model, however, we lack such identifiability results precisely

because the random effect and error term are combined linearly and their distributions cannot be distinguished.) As a result, if two sets of parameters give rise to the same marginal survival function, then they must do so by giving rise to the same random-effect distribution. Based on the proportional hazards structure, we prove a parallel result for the invertibility of the information operator: the existence of a submodel with zero score implies that the random-effect distribution has zero score along that submodel as well. Therefore, to ensure the invertibility of the information operator of the mixture Cox model, one only has to ensure that the information matrix of the random-effect distribution is invertible.

Our work can be extended in several directions. First, one may be interested in expanding the model by inclusion of more latent and observed variables. As the number of variables increases, the number of parameters to be estimated increases as well. Then, it may be desirable to perform variable selection. Because a single variable may be associated with multiple parameters, one may prefer not to treat parameters as the basic unit of selection, as in traditional lasso methods (Tibshirani 1996). Instead, methods like group lasso (Yuan and Lin 2006) that penalize parameters associated with a variable as a group may be considered.

In our model, the distribution of the manifest variable $\boldsymbol{Y}$ is fully parametric. One can allow a nonparametric transformation on $\boldsymbol{Y}$. A major challenge arises in extending the asymptotic results to unbounded nonparametric transformation, as the estimator of the transformation function can be unbounded (Zeng and Lin 2010).

Finally, it would be of interest to consider interval-censored data. Interval censoring results in a different likelihood function, which makes the computation of the NPMLE and the derivation of its asymptotic properties challenging, even for univariate survival time data. The asymptotic theory for interval-censored data is only available in a few simple cases; see Huang and Wellner (1997) for a review.

2.8 Technical Details

We present the following conditions, which are clearly implied by conditions (C3)-(C5):

(C3') For $k = 1, \ldots, K_1$, $\boldsymbol{\vartheta}_k$ is a non-zero vector.

(C4') Consider two sets of parameters $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\psi}, \boldsymbol{\nu})$ and $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ for $k = 1, \ldots, K_1$, and let $\tilde{\boldsymbol{\eta}}_1 = (\tilde{\eta}_{11}, \ldots, \tilde{\eta}_{1K_1})$, where $\tilde{\eta}_{1k} = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{\alpha}_k - \tilde{\boldsymbol{\alpha}}_k) + \boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + \eta_{1k}$. Let f_{Y,η_1} be the density

of $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$ given $\boldsymbol{Z}$. Then, $f_{Y,\eta_1}(\boldsymbol{Y}, \boldsymbol{\eta}_1 \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu}) = f_{Y,\eta_1}(\boldsymbol{Y}, \tilde{\boldsymbol{\eta}}_1 \mid \boldsymbol{Z}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ for all $\boldsymbol{Z}$, $\boldsymbol{Y}$, and $\boldsymbol{\eta}_1$ implies that $\boldsymbol{\alpha}_k = \tilde{\boldsymbol{\alpha}}_k$, $\boldsymbol{\beta}_k = \tilde{\boldsymbol{\beta}}_k$, $\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}$, and $\boldsymbol{\nu} = \tilde{\boldsymbol{\nu}}$.

(C5') For $k = K_1 + 1, \ldots, K$, let $(\boldsymbol{Y}^{(k)}, \boldsymbol{\eta}_1^{(k)}, \boldsymbol{\eta}_2^{(k)})$ be the components of $(\boldsymbol{Y}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ that appear in the regression of T_k, and let $(\boldsymbol{Y}^{-(k)}, \boldsymbol{\eta}_1^{-(k)}, \boldsymbol{\eta}_2^{-(k)})$ be the remaining components. If $\boldsymbol{\eta}_2^{(k)}$ is non-empty, then $(\boldsymbol{Y}^{-(k)}, \boldsymbol{\eta}_1^{-(k)})$ is non-empty, and $\boldsymbol{\eta}_2^{(k)}$ is complete sufficient in $\{F_{\boldsymbol{\eta}_2^{(k)}|Y,\eta_1}(\cdot \mid \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\eta}_1) : \boldsymbol{Z} = \boldsymbol{z}_0, \boldsymbol{Y}^{(k)} = \boldsymbol{y}_0, \boldsymbol{\eta}_1^{(k)} = \boldsymbol{\eta}_{10}\}$ for any fixed $\boldsymbol{z}_0$, $\boldsymbol{y}_0$, and $\boldsymbol{\eta}_{10}$, where $F_{\boldsymbol{\eta}_2^{(k)}|Y,\eta_1}$ is the distribution function of $\boldsymbol{\eta}_2^{(k)}$ given $(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\eta}_1)$ with $(\boldsymbol{Y}^{-(k)}, \boldsymbol{\eta}_1^{-(k)})$ treated as a parameter vector.

We prove Theorem 2.1 under the generalized conditions (C1), (C2), and (C3')-(C5'). The proof makes use of two lemmas given at the end of this section. We first provide an overview of the proof. For any two sets of parameters $(\boldsymbol{\vartheta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\phi}_k, \Lambda_k, \boldsymbol{\psi}, \boldsymbol{\nu})$ and $(\tilde{\boldsymbol{\vartheta}}_k, \tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\phi}}_k, \tilde{\Lambda}_k, \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$, assume that the likelihood values at the two sets of parameters are identical almost surely. By definition, the model is identifiable if the equality of the likelihood values implies the equality of the two sets of parameters. We derive the equality of the two sets of parameters in the following steps:

1. By conditions (C1), (C2), and (C3') and the identifiability of the mixture Cox model (Kortram et al. 1995), $\boldsymbol{\vartheta}_k = \tilde{\boldsymbol{\vartheta}}_k$ and $\Lambda_k = \tilde{\Lambda}_k$ for $k = 1, \ldots, K_1$.

2. With some algebraic manipulation, the likelihood function can be expressed in the form of the Laplace transform of the distribution of a function of $(\boldsymbol{Y}, \boldsymbol{\eta}_1)$. The uniqueness of the Laplace transform, together with condition (C4'), implies that $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\psi}, \boldsymbol{\nu}) = (\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ for $k = 1, \ldots, K_1$.

3. By the uniqueness of the Laplace transform and the complete sufficiency of $\boldsymbol{\eta}_2$ imposed by condition (C5'), the equality of the likelihood functions of $(T_{K_1+1}, \ldots, T_K, \boldsymbol{Y})$ implies the equality of the likelihood functions of $(T_{K_1+1}, \ldots, T_K, \boldsymbol{Y}, \boldsymbol{\eta})$. By the identifiability of the Cox model, we conclude that $(\boldsymbol{\vartheta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\phi}_k) = (\tilde{\boldsymbol{\vartheta}}_k, \tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\phi}}_k)$ for $k = K_1 + 1, \ldots, K$.

Proof of Theorem 2.1. The likelihood is given in (2.4). Here, we consider a single observation and drop the subscript i. Using the arguments in Section 10.1 of Zeng and Lin (2010), we can set each survival time to be right censored at any time point within $[0, \tau]$ when establishing identifiability.

Consider two sets of parameters $(\boldsymbol{\vartheta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\phi}_k, \Lambda_k, \boldsymbol{\psi}, \boldsymbol{\nu})$ and $(\tilde{\boldsymbol{\vartheta}}_k, \tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\phi}}_k, \tilde{\Lambda}_k, \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ such that the likelihood values for an observation with the K survival times being right censored are equal almost surely, i.e.,

$$\begin{aligned}
&\int \prod_{k=1}^{K} \left[\int \exp\left\{ -\Lambda_k\left(t_k\right) s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} \right\} g_k\left(s_k\right) \mathrm{d}m_k(s_k) \right] \\
&\qquad \times f_{Y,\eta}(\boldsymbol{Y}, \boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\eta} \\
&= \int \prod_{k=1}^{K} \left[\int \exp\left\{ -\tilde{\Lambda}_k\left(t_k\right) s_k e^{\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{\vartheta}}_k + \boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_k + \boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \boldsymbol{\eta}^{\mathrm{T}}\tilde{\boldsymbol{\phi}}_k} \right\} g_k\left(s_k\right) \mathrm{d}m_k(s_k) \right] \\
&\qquad \times f_{Y,\eta}(\boldsymbol{Y}, \boldsymbol{\eta} \mid \boldsymbol{Z}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}}) \, \mathrm{d}\boldsymbol{\eta}
\end{aligned} \tag{2.5}$$

for all $t_1, \ldots, t_K \in [0, \tau]$, $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$, where $f_{Y,\eta}$ is the density of $(\boldsymbol{Y}, \boldsymbol{\eta})$ given $\boldsymbol{Z}$. If m_k is a point mass at one, then s_k is fixed at one, $g_k = 1$, and the integration with respect to $m_k(s_k)$ can be omitted. For simplicity of description, assume that m_k is the Lebesgue measure. Note that

$$\int s g_k\left(s\right) \, \mathrm{d}s = -\lim_{t \to 0^+} \frac{\mathrm{d}}{\mathrm{d}t} \int_0^\infty e^{-ts} g_k\left(s\right) \, \mathrm{d}s = -\lim_{t \to 0^+} \frac{\mathrm{d}}{\mathrm{d}t} \exp\left\{-G_k\left(t\right)\right\} = G_k'\left(0\right) < \infty.$$

Thus, a transformation model can be written as a random-effect proportional hazards model with known distributions $(g_1, \ldots, g_K)$ for random effects $(s_1, \ldots, s_K)$ with finite means.

First, we show that the baseline hazard functions of the first K_1 survival times are identifiable. For each $k = 1, \ldots, K_1$, set $t_l \to 0$ for $l \neq k$ on both sides of (2.5). On each side of the resulting equation, integration with respect to $\boldsymbol{Y}$ results in the likelihood of a mixture Cox model with $s_k e^{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \eta_{1k}}$ or $s_k e^{\boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \eta_{1k}}$ as a latent variable. Let $\mathrm{E}(\cdot \mid \boldsymbol{Z})$ and $\tilde{\mathrm{E}}(\cdot \mid \boldsymbol{Z})$ be the expectations under $f_{Y,\eta}(\cdot \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu})$ and $f_{Y,\eta}(\cdot \mid \boldsymbol{Z}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$, respectively. Theorem 3 of Kortram et al. (1995) implies that $\mathrm{E}(s_k e^{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \eta_{1k}} \mid \boldsymbol{Z} = \boldsymbol{0})\Lambda_k = \tilde{\mathrm{E}}(s_k e^{\boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \eta_{1k}} \mid \boldsymbol{Z} = \boldsymbol{0})\tilde{\Lambda}_k$ on $[0, \tau]$, $\boldsymbol{\vartheta}_k = \tilde{\boldsymbol{\vartheta}}_k$, and the distribution of $\mathrm{E}(s_k e^{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \eta_{1k}} \mid \boldsymbol{Z} = \boldsymbol{0})^{-1} s_k e^{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \eta_{1k}}$ under $f_{Y,\eta}(\cdot \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu})$ is equal to that of $\tilde{\mathrm{E}}(s_k e^{\boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \eta_{1k}} \mid \boldsymbol{Z} = \boldsymbol{0})^{-1} s_k e^{\boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \eta_{1k}}$ under $f_{Y,\eta}(\cdot \mid \boldsymbol{Z}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$. Because $\mathrm{E}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \eta_{1k} \mid \boldsymbol{Z} = \boldsymbol{0}) = \tilde{\mathrm{E}}(\boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \eta_{1k} \mid \boldsymbol{Z} = \boldsymbol{0}) = 0$ by condition (C2), we see that $\Lambda_k = \tilde{\Lambda}_k$ on $[0, \tau]$.

Second, we show that the likelihood function takes the form of a Laplace transform and use the uniqueness of the Laplace transform to prove the identifiability of $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\psi}, \boldsymbol{\nu})$ $(k = 1, \ldots, K_1)$. Setting $t_k \to 0$ for $k = K_1 + 1, \ldots, K$ and $\boldsymbol{W} = \boldsymbol{0}$ on both sides of (2.5), we have

$$\int \prod_{k=1}^{K_1} \left[\int \exp\left\{ -\Lambda_k\left(t_k\right) s_k e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \eta_{1k}} \right\} g_k\left(s_k\right) \mathrm{d}s_k \right] f_{Y,\eta}\left(\boldsymbol{Y}, \boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu}\right) \mathrm{d}\boldsymbol{\eta}$$

$$= \int \prod_{k=1}^{K_1} \left[\int \exp\left\{ -\Lambda_k\left(t_k\right) s_k e^{\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_k + \boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \eta_{1k}} \right\} g_k\left(s_k\right) \mathrm{d}s_k \right] f_{Y,\eta}\left(\boldsymbol{Y}, \boldsymbol{\eta} \mid \boldsymbol{Z}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}}\right) \mathrm{d}\boldsymbol{\eta}. \tag{2.6}$$

Let $\boldsymbol{U} = (U_1, \ldots, U_{K_1})$, $U_k = s_k e^{\eta_{1k}}$, and $f_{U|Y}$ be the density function of $\boldsymbol{U}$ given $\boldsymbol{Z}$ and $\boldsymbol{Y}$. By the uniqueness of the Laplace transform, for any continuous functions f and $\tilde{f}$, any open set $\mathcal{S}$, and any positive real numbers c and $\tilde{c}$,

$$\int_0^\infty e^{-cst} f\left(t\right) \mathrm{d}t = \int_0^\infty e^{-\tilde{c}st} \tilde{f}\left(t\right) \mathrm{d}t \quad \forall s \in \mathcal{S}$$

implies that $f(t) = (c/\tilde{c})\tilde{f}(ct/\tilde{c})$ for all $t > 0$. Therefore, the equality of (2.6) for all $t_1, \ldots, t_{K_1}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$ implies that

$$f_{U|Y}(\boldsymbol{U} \mid \boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{\psi}, \boldsymbol{\nu}) = e^{\sum_{k=1}^{K_1} \boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{\alpha}_k - \tilde{\boldsymbol{\alpha}}_k)} f_{U|Y}(\tilde{\boldsymbol{U}} \mid \boldsymbol{Z}, \boldsymbol{Y}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}}) \tag{2.7}$$

for all $\boldsymbol{U}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$, where $\tilde{\boldsymbol{U}} = (\tilde{U}_1, \ldots, \tilde{U}_{K_1})$, and $\tilde{U}_k = e^{\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{\alpha}_k - \tilde{\boldsymbol{\alpha}}_k)} U_k$. Let $f_{\eta_1|Y}$ be the density of $\boldsymbol{\eta}_1$ given $\boldsymbol{Z}$ and $\boldsymbol{Y}$. By the definition of $\boldsymbol{U}$,

$$\begin{aligned}
&f_{U|Y}\left(\boldsymbol{U} \mid \boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{\psi}, \boldsymbol{\nu}\right) \\
&= \int_{s_k > 0} f_{\eta_1|Y}\left(\log U_1 - \log s_1, \ldots, \log U_{K_1} - \log s_{K_1} \mid \boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{\psi}, \boldsymbol{\nu}\right) \prod_{k=1}^{K_1} U_k^{-1} g_k\left(s_k\right) \mathrm{d}\left(s_1, \ldots, s_{K_1}\right) \\
&= \int_{\mathbb{R}^{K_1}} f_{\eta_1|Y}\left(\log U_1 - v_1, \ldots, \log U_{K_1} - v_{K_1} \mid \boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{\psi}, \boldsymbol{\nu}\right) \prod_{k=1}^{K_1} U_k^{-1} \bar{g}_k\left(v_k\right) e^{v_k} \mathrm{d}\left(v_1, \ldots, v_{K_1}\right),
\end{aligned}$$

where $\bar{g}_k(v) = g_k(e^v)$. Thus, (2.8) implies that

$$\begin{aligned}
\int_{\mathbb{R}^{K_1}} \Big\{ & f_{\eta_1|Y}\big(\log U_1 - v_1, \ldots, \log U_{K_1} - v_{K_1} \mid \boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{\psi}, \boldsymbol{\nu}\big) \\
- & f_{\eta_1|Y}\big(\log \tilde{U}_1 - v_1, \ldots, \log \tilde{U}_{K_1} - v_{K_1} \mid \boldsymbol{Z}, \boldsymbol{Y}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}}\big) \Big\} \prod_{k=1}^{K_1} \bar{g}_k\left(v_k\right) e^{v_k} \mathrm{d}\left(v_1, \ldots, v_{K_1}\right) = 0.
\end{aligned} \tag{2.8}$$

Consider two arbitrary continuous functions $f, g : \mathbb{R} \to \mathbb{R}$. Note that

$$\int_{-\infty}^{\infty} e^{-st} (f * g)(t) \, \mathrm{d}t = \int_{-\infty}^{\infty} e^{-st} f(t) \, \mathrm{d}t \int_{-\infty}^{\infty} e^{-st} g(t) \, \mathrm{d}t$$

for any s such that the integrals are defined, where $(f*g)(t) \equiv \int_{-\infty}^{\infty} f(t-s)g(s) \, \mathrm{d}s$ is the convolution of f and g. Therefore, $(f * g)(\cdot) = 0$ implies that

$$\int_{-\infty}^{\infty} e^{-st} f(t) \, \mathrm{d}t \int_{-\infty}^{\infty} e^{-st} g(t) \, \mathrm{d}t = 0,$$

which, if g is positive, implies that $f(\cdot) = 0$ by the uniqueness of the bilateral Laplace transform (Chareka 2007). Because $\bar{g}_k(\cdot)e^{(\cdot)}$ is positive, (2.8) implies that $f_{\eta_1|Y}(\boldsymbol{\eta}_1 \mid \boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{\psi}, \boldsymbol{\nu}) = f_{\eta_1|Y}(\tilde{\boldsymbol{\eta}}_1 \mid \boldsymbol{Z}, \boldsymbol{Y}; \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$, where $\tilde{\boldsymbol{\eta}}_1$ is defined in condition (C4'). By condition (C4'), $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\psi}, \boldsymbol{\nu}) = (\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ for $k = 1, \ldots, K_1$.

It remains to identify the parameters associated with $(T_{K_1+1}, \ldots, T_K)$. By the uniqueness of the Laplace transform, (2.5) implies that

$$\begin{aligned}
&\int \prod_{k=K_1+1}^{K} \left\{ \int e^{-\Lambda_k(t_k) s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}} g_k(s_k) \, \mathrm{d}s_k \right\} f_{Y,\eta}(\boldsymbol{Y}, \boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\eta}_2 \\
=&\int \prod_{k=K_1+1}^{K} \left\{ \int e^{-\tilde{\Lambda}_k(t_k) s_k e^{\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{\vartheta}}_k + \boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_k + \boldsymbol{Y}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}_k + \boldsymbol{\eta}^{\mathrm{T}}\tilde{\boldsymbol{\phi}}_k}} g_k(s_k) \, \mathrm{d}s_k \right\} f_{Y,\eta}(\boldsymbol{Y}, \boldsymbol{\eta} \mid \boldsymbol{Z}; \boldsymbol{\psi}, \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\eta}_2
\end{aligned}$$

for all $t_{K_1+1}, \ldots, t_K$, $\boldsymbol{W}$, $\boldsymbol{Z}$, $\boldsymbol{Y}$, and $\boldsymbol{\eta}_1$, i.e., $\boldsymbol{\eta}_1$ can be treated as observed for identifying the remaining parameters. Under condition (C5'), we can use the arguments in the proof of Lemma 2.1 to show that the integrands in the above equality are equal at each value of $\boldsymbol{\eta}_2$. We conclude that $(\boldsymbol{\vartheta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\phi}_k, \Lambda_k) = (\tilde{\boldsymbol{\vartheta}}_k, \tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\phi}}_k, \tilde{\Lambda}_k)$ for $k = K_1 + 1, \ldots, K$. □

We provide an overview for the proof of Theorem 2.2. The consistency of the NPMLE is proved in the following steps:

1. By conditions (D2)-(D4), the NPMLE exists, i.e., $\hat{\Lambda}_k(\tau) < \infty$.

2. By conditions (D3) and (D4), $\hat{\Lambda}_k(\tau)$ is uniformly bounded. Helly's selection theorem then implies that every subsequence of $\hat{\Lambda}_k$ has a further converging subsequence.

3. By the Glivenko-Cantelli properties of the log-likelihood and related functions given by Lemma 2.3, the identifiability of the model, and the non-negativity of the Kullback-Leibler divergence,

we conclude the consistency of the NPMLE.

The asymptotic normality of the NPMLE follows mainly from the arguments of van der Vaart (1998, pp. 419–424). Donsker properties of the score and related functions are given by Lemma 2.4, and the invertibility of the information operator is given by Lemma 2.2.

Proof of Theorem 2.2. We use $\boldsymbol{Z}$ to denote both $\boldsymbol{W}$ and $\boldsymbol{Z}$ with $\boldsymbol{\beta}_k$ $(k = 1, \ldots, K)$ being the corresponding vector of regression parameters. Let

$$\begin{aligned}\Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A}) = \prod_{k=1}^{K} \int \Big[e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} G_k' \left\{e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} \Lambda_k(\tilde{T}_{ki})\right\}\Big]^{\Delta_{ki}} \\ \times \exp\left[-G_k\left\{e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} \Lambda_k(\tilde{T}_{ki})\right\}\right] f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\eta}; \boldsymbol{\psi}) f_\eta(\boldsymbol{\eta} \mid \boldsymbol{Z}_i; \boldsymbol{\nu})\, \mathrm{d}\boldsymbol{\eta},\end{aligned}$$

$\dot{\Psi}_\theta(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})$ be the derivative of $\Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})$ with respect to $\boldsymbol{\theta}$, and $\dot{\Psi}_k(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})[H_k]$ be the derivative of $\Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})$ along the path $(\Lambda_k + \epsilon H_k)$.

First, we prove the consistency. By condition (D4),

$$\begin{aligned}&\Big[e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} G_k' \left\{e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} \Lambda_k(\tilde{T}_{ki})\right\}\Big]^{\Delta_{ki}} e^{-G_k\left\{e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} \Lambda_k(\tilde{T}_{ki})\right\}} \\ \leq\ & e^{O(1+|\boldsymbol{Y}|+|\boldsymbol{\eta}|)}\big\{1 + \Lambda_k(\tilde{T}_{ki})\big\}^{-\Delta_{ki} - \kappa_{3k} + \kappa_{2k} + 1}.\end{aligned}$$

Thus, condition (D3) implies that

$$\Psi\left(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A}\right) \leq \prod_{i=1}^{n} \mathcal{F}(\mathcal{O}_i; \boldsymbol{\theta}) \prod_{k=1}^{K} \left\{1 + \Lambda_k(\tilde{T}_{ki})\right\}^{-\Delta_{ki} - \kappa_{3k} + \kappa_{2k} + 1}, \tag{2.9}$$

where $\mathcal{F}(\mathcal{O}_i; \boldsymbol{\theta})$ is a random variable with $|\mathrm{E}\{\log \mathcal{F}(\mathcal{O}_i; \boldsymbol{\theta})\}| < \infty$ for any $\boldsymbol{\theta}$. By condition (D2), $P(\tilde{T}_{ki} = \tau)$ is positive. Therefore, if $\Lambda_k(\tau) = \infty$, then the right-hand side of (2.9) is zero for large n. We conclude that $\hat{\Lambda}_k(\tau) < \infty$, such that the NPMLE exists.

We then show that $\limsup_n \hat{\Lambda}_k(\tau) < \infty$ almost surely. From (2.9),

$$\begin{aligned}\frac{1}{n} \log L_n(\hat{\boldsymbol{\theta}}, \hat{\mathcal{A}}) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ki} \log \hat{\Lambda}_k\{\tilde{T}_{ki}\} + \frac{1}{n} \sum_{i=1}^{n} \log \Psi(\mathcal{O}_i; \hat{\boldsymbol{\theta}}, \hat{\mathcal{A}}) \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \log \mathcal{F}(\mathcal{O}_i; \hat{\boldsymbol{\theta}}) + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ki} \log \hat{\Lambda}_k\{\tilde{T}_{ki}\}\end{aligned}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\Delta_{ki}+\kappa_{3k}-\kappa_{2k}-1\right)\log\left\{1+\hat{\Lambda}_k(\tilde{T}_{ki})\right\}.$$

Let $\tilde{N}=n^{-1}\sum_{i=1}^{n}(\Delta_{1i}I(\tilde{T}_{1i}\leq\cdot),\ldots,\Delta_{Ki}I(\tilde{T}_{Ki}\leq\cdot))$. Clearly,

$$\frac{1}{n}\log L_n(\boldsymbol{\theta}_0,\tilde{N})=-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ki}\log n+\frac{1}{n}\sum_{i=1}^{n}\log\Psi(\mathcal{O}_i;\boldsymbol{\theta}_0,\tilde{N}).$$

The second term on the right-hand side of the above equation is $O_p\left(1\right)$. Thus,

$$\begin{aligned}&\frac{1}{n}\log L_n(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})-\frac{1}{n}\log L_n(\boldsymbol{\theta}_0,\tilde{N})+O_p(1)\\ \leq\ &\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ki}\log\left[n\hat{\Lambda}_k\{\tilde{T}_{ki}\}\right]-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\Delta_{ki}+\kappa_{3k}-\kappa_{2k}-1\right)\log\left\{1+\hat{\Lambda}_k(\tilde{T}_{ki})\right\}.\end{aligned}$$

Note that $(\kappa_{3k}-\kappa_{2k}-1)$ is positive by condition (D4). Using the partitioning argument similar to those of Murphy (1994) and Parner (1998), we can show that the right-hand side of the above inequality tends to $-\infty$ if $\limsup_n\hat{\Lambda}_k\left(\tau\right)=\infty$. By definition of $(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})$, the left-hand side of the inequality is bounded below by an $O_p(1)$ term. Therefore, $\hat{\Lambda}_k\left(\tau\right)$ is uniformly bounded.

Given the boundedness of $\hat{\Lambda}_k\left(\tau\right)$, Helly's selection theorem implies that, for any subsequence of n, we can always choose a further subsequence such that $\hat{\Lambda}_k$ converges pointwise to some monotone function Λ_k^* and $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$. The desired consistency result follows if we can show that $\Lambda_k^*=\Lambda_{0k}$ and $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ almost surely. With an abuse of notation, let $\{n\}_{1,2,\ldots}$ be the subsequence. Define

$$\tilde{\Lambda}_k\left(t\right)=-\sum_{i=1}^{n}\Delta_{ki}I(\tilde{T}_{ki}\leq t)\left\{\sum_{j=1}^{n}\frac{\dot{\Psi}_k(\mathcal{O}_j;\boldsymbol{\theta}_0,\mathcal{A}_0)[I(\tilde{T}_{ki}\leq\cdot)]}{\Psi(\mathcal{O}_j;\boldsymbol{\theta}_0,\mathcal{A}_0)}\right\}^{-1}.$$

By Lemma 2.4 and the properties of Donsker (and therefore, Glivenko-Cantelli) classes,

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\dot{\Psi}_k(\mathcal{O}_j;\boldsymbol{\theta}_0,\mathcal{A}_0)[I\left(s\leq\cdot\right)]}{\Psi(\mathcal{O}_j;\boldsymbol{\theta}_0,\mathcal{A}_0)}\rightarrow\mathrm{E}\left(\frac{\dot{\Psi}_k(\mathcal{O}_i;\boldsymbol{\theta}_0,\mathcal{A}_0)[I\left(s\leq\cdot\right)]}{\Psi(\mathcal{O}_i;\boldsymbol{\theta}_0,\mathcal{A}_0)}\right)$$

uniformly on $[0,\tau]$. Because the score function along the path $\Lambda_k=\Lambda_{0k}+\epsilon I\left(\cdot\geq s\right)$ with other parameters fixed at their true values has zero expectation,

$$-\mathrm{E}\left\{\frac{\dot{\Psi}_k(\mathcal{O}_i;\boldsymbol{\theta}_0,\mathcal{A}_0)[I\left(s\leq\cdot\right)]}{\Psi(\mathcal{O}_i;\boldsymbol{\theta}_0,\mathcal{A}_0)}\right\}=\frac{\mathrm{d}P(\tilde{T}_{ki}\Delta_{ki}<s)/\mathrm{d}s}{\lambda_{0k}\left(s\right)}.$$

Algebraic manipulation yields that the uniform limit of $\tilde{\Lambda}_k$ on $[0,\tau]$ is Λ_{0k}. Note that

$$\hat{\Lambda}_k(t) = \int_0^t \frac{\left| n^{-1}\sum_{j=1}^n \dot{\Psi}_k(\mathcal{O}_j;\boldsymbol{\theta}_0,\mathcal{A}_0)[I(s\le\cdot)]/\Psi(\mathcal{O}_j;\boldsymbol{\theta}_0,\mathcal{A}_0)\right|}{\left| n^{-1}\sum_{j=1}^n \dot{\Psi}_k(\mathcal{O}_j;\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})[I(s\le\cdot)]/\Psi(\mathcal{O}_j;\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})\right|} \mathrm{d}\tilde{\Lambda}_k(s).$$

We have shown that the numerator of the integrand in the above equation converges uniformly. Similarly, we can show that the denominator of the integrand in the above equation converges uniformly to $|\mathrm{E}\{\dot{\Psi}_k(\mathcal{O}_i;\boldsymbol{\theta}^*,\mathcal{A}^*)[I(s\le\cdot)]/\Psi(\mathcal{O}_i;\boldsymbol{\theta}^*,\mathcal{A}^*)\}|$ and that the limit is bounded away from 0. Because $\tilde{\Lambda}_k$ converges uniformly to Λ_{0k}, which is differentiable with respect to t, Λ_k^* is also differentiable with respect to t. It follows that $\mathrm{d}\hat{\Lambda}_k/\mathrm{d}\tilde{\Lambda}_k$ converges uniformly to λ_k^*/λ_{0k} on $[0,\tau]$, where $\lambda_k^* = (\Lambda_k^*)'$. As $n^{-1}\log L_n(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}}) - n^{-1}\log L_n(\boldsymbol{\theta}_0,\tilde{\mathcal{A}})$ is non-negative,

$$\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \Delta_{ki}\log\frac{\mathrm{d}\hat{\Lambda}_k(\tilde{T}_{ki})}{\mathrm{d}\tilde{\Lambda}_k(\tilde{T}_{ki})} + \frac{1}{n}\sum_{i=1}^n \log\frac{\Psi(\mathcal{O}_i;\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})}{\Psi(\mathcal{O}_i;\boldsymbol{\theta}_0,\tilde{\mathcal{A}})} \ge 0.$$

By the Glivenko-Cantelli properties of the class of functions of $\log\Psi(\mathcal{O}_i;\boldsymbol{\theta},\mathcal{A})$ given by Lemma 2.4 and the uniform convergence of $\mathrm{d}\hat{\Lambda}_k/\mathrm{d}\tilde{\Lambda}_k$, setting $n\to\infty$ on both sides of the above inequality yields

$$\mathrm{E}\left\{\log\frac{\prod_{k=1}^K \lambda_k^*(\tilde{T}_{ki})^{\Delta_{ki}}\Psi(\mathcal{O}_i;\boldsymbol{\theta}^*,\mathcal{A}^*)}{\prod_{k=1}^K \lambda_{0k}(\tilde{T}_{ki})^{\Delta_{ki}}\Psi(\mathcal{O}_i;\boldsymbol{\theta}_0,\mathcal{A}_0)}\right\} \ge 0.$$

The left-hand side of the above inequality is the negative Kullback-Leibler distance of the density indexed by $(\boldsymbol{\theta}^*,\mathcal{A}^*)$. From the identifiability of the model implied by Theorem 2.1, we conclude that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda_k^* = \Lambda_{k0}$. The desired consistency result follows.

To prove the asymptotic normality of the NPMLE, we adopt the arguments of van der Vaart (1998, pp. 419–424). Let $\mathcal{P}_n$ be the empirical measure determined by n i.i.d. observations, and let $\mathcal{P}$ be the true probability measure. Let $\dot{\ell}_\theta(\boldsymbol{\theta},\mathcal{A})$ be the derivative of $\log L_n(\boldsymbol{\theta},\mathcal{A})$ with respect to $\boldsymbol{\theta}$, and let $\dot{\ell}_k(\boldsymbol{\theta},\mathcal{A})[H_k]$ be the derivative of $\log L_n(\boldsymbol{\theta},\mathcal{A})$ along the path $(\Lambda_k + \epsilon H_k)$. For any $\boldsymbol{v}\in\mathbb{R}^d$ and $\mathcal{W} = (h_1,\ldots,h_K)$ with $h_k \in BV[0,\tau]$, where $BV[0,\tau]$ is the space of functions of bounded variation on $[0,\tau]$, we have

$$\mathcal{P}_n\left(\boldsymbol{v}^{\mathrm{T}}\dot{\ell}_\theta(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}}) + \sum_{k=1}^K \dot{\ell}_k(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})[\int h_k\,\mathrm{d}\hat{\Lambda}_k]\right) = 0.$$

In addition,

$$\mathcal{P}\Big(\boldsymbol{v}^{\mathrm{T}}\dot{\ell}_{\theta}(\boldsymbol{\theta}_0,\mathcal{A}_0)+\sum_{k=1}^{K}\dot{\ell}_k(\boldsymbol{\theta}_0,\mathcal{A}_0)[\int h_k\,\mathrm{d}\Lambda_{0k}]\Big)=0.$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(\mathcal{P}_n-\mathcal{P})\Big(\boldsymbol{v}^{\mathrm{T}}\dot{\ell}_{\theta}(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})+\sum_{k=1}^{K}\dot{\ell}_k(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})[\int h_k\,\mathrm{d}\hat{\Lambda}_k]\Big) \qquad (2.10)\\
= \;& -\sqrt{n}\mathcal{P}\Big\{\Big(\boldsymbol{v}^{\mathrm{T}}\dot{\ell}_{\theta}(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})+\sum_{k=1}^{K}\dot{\ell}_k(\hat{\boldsymbol{\theta}},\hat{\mathcal{A}})[\int h_k\,\mathrm{d}\hat{\Lambda}_k]\Big)\\
&-\Big(\boldsymbol{v}^{\mathrm{T}}\dot{\ell}_{\theta}(\boldsymbol{\theta}_0,\mathcal{A}_0)+\sum_{k=1}^{K}\dot{\ell}_k(\boldsymbol{\theta}_0,\mathcal{A}_0)[\int h_k\,\mathrm{d}\Lambda_{0k}]\Big)\Big\}.
\end{aligned}$$

From the Donsker properties of the classes of functions of $\dot{\ell}_{\theta}$ and $\dot{\ell}_k$ implied by Lemma 2.4 and the consistency of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathcal{A}}$, we conclude that the left-hand side of (2.10) equals

$$\sqrt{n}(\mathcal{P}_n-\mathcal{P})\Big(\boldsymbol{v}^{\mathrm{T}}\dot{\ell}_{\theta}(\boldsymbol{\theta}_0,\mathcal{A}_0)+\sum_{k=1}^{K}\dot{\ell}_k(\boldsymbol{\theta}_0,\mathcal{A}_0)[\int h_k\,\mathrm{d}\Lambda_{0k}]\Big)+o_p\left(1\right).$$

This term converges to a Gaussian process in $l^{\infty}(\mathcal{V}\times\mathcal{Q}^K)$. By the Taylor series expansion, the right-hand side of (2.10) is of the form

$$\begin{aligned}
&-\sqrt{n}\left\{B_1\left[\boldsymbol{v},\mathcal{W}\right]^{\mathrm{T}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+\sum_{k=1}^{K}\int B_{2k}\left[\boldsymbol{v},\mathcal{W}\right]\,\mathrm{d}(\hat{\Lambda}_k-\Lambda_{0k})\right\}\\
&+o_p\Big(\sqrt{n}\left|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right|+\sqrt{n}\sum_{k=1}^{K}\left\|\hat{\Lambda}_k-\Lambda_{0k}\right\|_{V[0,\tau]}\Big),
\end{aligned}$$

where $\mathcal{B}\equiv(B_1,B_{21},\ldots,B_{2K})$ is the information operator and is linear in $\mathbb{R}^d\times BV\left[0,\tau\right]^K$. By Lemma 2.2, $\mathcal{B}$ is invertible. The rest of the proof then follows the arguments of van der Vaart (1998, pp. 419–424). Finally, because $\boldsymbol{v}^{\mathrm{T}}\hat{\boldsymbol{\theta}}$ is an asymptotically linear estimator of $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\theta}_0$ with the influence function lying in the space spanned by the score functions, $\hat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$. □

The following four lemmas are used to prove Theorem 2.1 and Theorem 2.2.

Lemma 2.1. *Let Model A be*

$$\begin{aligned} \boldsymbol{X} \mid (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \stackrel{d}{=} \boldsymbol{X} \mid \boldsymbol{\eta}_1 &\sim F_{X|\eta_1}(\cdot \mid \boldsymbol{\eta}_1), \\ \boldsymbol{Y} \mid (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) &\sim F_{Y|\eta}(\cdot \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \\ \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1 &\sim F_{\eta_2|\eta_1}(\cdot \mid \boldsymbol{\eta}_1), \\ \boldsymbol{\eta}_1 &\sim F_{\eta_1}, \end{aligned}$$

where $(\boldsymbol{X}, \boldsymbol{Y})$ *are observed, and* $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ *are latent. Model A is depicted in Figure 2.5. Let* $f_{X|\eta_1} = F'_{X|\eta_1}$, $f_{Y|\eta} = F'_{Y|\eta}$, $f_{\eta_2|\eta_1} = F'_{\eta_2|\eta_1}$, *and* $f_{\eta_1} = F'_{\eta_1}$. *Assume that: (a) for any density functions* $\tilde{f}_{Y|\eta}$ *and* $\tilde{f}_{\eta_2|\eta_1}$,

$$\int f_{Y|\eta}(\boldsymbol{Y} \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) f_{\eta_2|\eta_1}(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1) \, \mathrm{d}\boldsymbol{\eta}_2 = \int \tilde{f}_{Y|\eta}(\boldsymbol{Y} \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \tilde{f}_{\eta_2|\eta_1}(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1) \, \mathrm{d}\boldsymbol{\eta}_2 \quad \forall \boldsymbol{Y}, \boldsymbol{\eta}_1$$

implies that $(f_{Y|\eta}, f_{\eta_2|\eta_1}) = (\tilde{f}_{Y|\eta}, \tilde{f}_{\eta_2|\eta_1})$, *i.e., the model for* $\boldsymbol{Y}$ *is identifiable if* $\boldsymbol{\eta}_1$ *is observed; (b)* $F_{X|\eta_1}$ *and* F_{η_1} *are identifiable based on* $(\boldsymbol{X}, \boldsymbol{Y})$*; and (c)* $\boldsymbol{\eta}_1$ *is a complete sufficient statistic in* $\{F_{\eta_1|X}(\cdot \mid \boldsymbol{X}) : \boldsymbol{X} \in \mathcal{X}\}$, *where* $F_{\eta_1|X}$ *is the conditional distribution function of* $\boldsymbol{\eta}_1$ *given* $\boldsymbol{X}$, *and* $\mathcal{X}$ *is the range of* $\boldsymbol{X}$. *Then, Model A is identifiable. A sufficient condition for* $\boldsymbol{\eta}_1$ *to be complete sufficient is that the density of* $\boldsymbol{X}$ *is of the form*

$$f_{X|\eta_1}(\boldsymbol{X} \mid \boldsymbol{\eta}_1) \propto \prod_{j=1}^{q} \exp\{X_j s_j(\boldsymbol{\eta}_1) - a_j(\boldsymbol{\eta}_1)\} b_j(X_j),$$

where $\boldsymbol{X} = (X_1, \ldots, X_q)$, $\boldsymbol{\eta}_1 \mapsto (s_1(\boldsymbol{\eta}_1), \ldots, s_q(\boldsymbol{\eta}_1))$ *is one-to-one, and* b_j *is non-zero on some open set.*

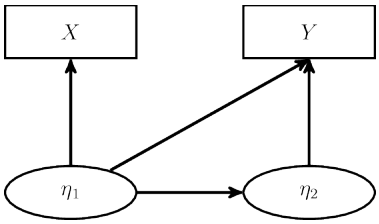


Figure 2.5. SEM Considered in Lemma 2.1. The SEM consists of two sets of latent variables and two sets of observed variables that may all be multivariate. The observed variable $\boldsymbol{X}$ *depends only on the latent variable* $\boldsymbol{\eta}_1$, *but the observed variable* $\boldsymbol{Y}$ *depends on both sets of latent variables.*

Lemma 2.2. *Under conditions (C1), (C2), (C3'), (C5'), and (D5), the model given by* (2.1)-(2.3) *has an invertible information operator.*

Lemma 2.3. *Consider functions h and g in the space $l^\infty(\mathbb{R}^r \times \mathbb{R}^q)$. Assume that for any $\boldsymbol{Y} \in \mathbb{R}^r$ and $\boldsymbol{b} \in \mathbb{R}^q$, there exist $M_j > 0$, $c_j > 0$, and $\boldsymbol{N}_j \in \mathbb{R}^r$ $(j = 1, \ldots, q)$ such that*

$$\prod_{j=1}^{q} \exp(-M_j \left|b_j\right| + \boldsymbol{N}_j^{\mathrm{T}} \boldsymbol{Y} b_j - c_j b_j^2) \leq h\left(\boldsymbol{Y}, \boldsymbol{b}\right) \leq \prod_{j=1}^{q} \exp(M_j \left|b_j\right| + \boldsymbol{N}_j^{\mathrm{T}} \boldsymbol{Y} b_j - c_j b_j^2),$$

and there exists $K > 0$ such that $g(\boldsymbol{Y}, \boldsymbol{b}) \leq \exp\{K(1 + |\boldsymbol{Y}| + |\boldsymbol{b}|)\}$. Then, with $\mathbf{1}_q$ being a q-vector of ones, both

$$\frac{\int \exp\left(-e^{A_1 + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{Y} + \mathbf{1}_q^{\mathrm{T}} \boldsymbol{b}}\right) h\left(\boldsymbol{Y}, \boldsymbol{b}\right) g\left(\boldsymbol{Y}, \boldsymbol{b}\right) \mathrm{d}\boldsymbol{b}}{\int \exp\left(-e^{A_2 + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{Y} + \mathbf{1}_q^{\mathrm{T}} \boldsymbol{b}}\right) h\left(\boldsymbol{Y}, \boldsymbol{b}\right) \mathrm{d}\boldsymbol{b}} \tag{2.11}$$

and

$$\frac{\int h\left(\boldsymbol{Y}, \boldsymbol{b}\right) g\left(\boldsymbol{Y}, \boldsymbol{b}\right) \mathrm{d}\boldsymbol{b}}{\int h\left(\boldsymbol{Y}, \boldsymbol{b}\right) \mathrm{d}\boldsymbol{b}} \tag{2.12}$$

are bounded by $e^{O(1+|\boldsymbol{Y}|)}$ for any $A_1 \in \mathbb{R}$, $A_2 \in \mathbb{R}$, and $\boldsymbol{B} \in \mathbb{R}^r$.

Lemma 2.4. *The following classes are Donsker:*

$$\begin{aligned}
\mathcal{C}_1 &= \left\{\log \Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A}) : \|\Lambda_m\|_{V[0,\tau]} \leq c, m = 1, \ldots, K, \boldsymbol{\theta} \in \Theta\right\}, \\
\mathcal{C}_2 &= \left\{\frac{\dot{\Psi}_\theta(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})}{\Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})} : \|\Lambda_m\|_{V[0,\tau]} \leq c, m = 1, \ldots, K, \boldsymbol{\theta} \in \Theta\right\},
\end{aligned}$$

and

$$\mathcal{C}_{3k} = \left\{\frac{\dot{\Psi}_k(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})\left[H\right]}{\Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})} : \|\Lambda_m\|_{V[0,\tau]} \leq c, m = 1, \ldots, K, \boldsymbol{\theta} \in \Theta, \|H\|_{V[0,\tau]} \leq c\right\}$$

for $k = 1, \ldots, K$ and any fixed $c > 0$, where Ψ, $\dot{\Psi}_\theta$, and $\dot{\Psi}_k$ are defined in the proof of Theorem 2.2.

Proof of Lemma 2.1. Assume that there exist two sets of density functions $(f_{Y|\eta}, f_{\eta_2|\eta_1})$ and $(\tilde{f}_{Y|\eta}, \tilde{f}_{\eta_2|\eta_1})$, such that the marginal densities are identical for all $\boldsymbol{X}$ and $\boldsymbol{Y}$. That is,

$$\begin{aligned}
&\int f_{X|\eta_1}\left(\boldsymbol{X} \mid \boldsymbol{\eta}_1\right) f_{Y|\eta}\left(\boldsymbol{Y} \mid \boldsymbol{\eta}\right) f_{\eta_2|\eta_1}\left(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1\right) f_{\eta_1}\left(\boldsymbol{\eta}_1\right) \mathrm{d}\boldsymbol{\eta} \\
= &\int f_{X|\eta_1}\left(\boldsymbol{X} \mid \boldsymbol{\eta}_1\right) \tilde{f}_{Y|\eta}\left(\boldsymbol{Y} \mid \boldsymbol{\eta}\right) \tilde{f}_{\eta_2|\eta_1}\left(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1\right) f_{\eta_1}\left(\boldsymbol{\eta}_1\right) \mathrm{d}\boldsymbol{\eta},
\end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. Thus, with $f_{\eta_1|X} = F'_{\eta_1|X}$,

$$\int \left\{ \int f_{Y|\eta}\left(\boldsymbol{Y} \mid \boldsymbol{\eta}\right) f_{\eta_2|\eta_1}\left(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1\right) - \tilde{f}_{Y|\eta}\left(\boldsymbol{Y} \mid \boldsymbol{\eta}\right) \tilde{f}_{\eta_2|\eta_1}\left(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1\right) \mathrm{d}\boldsymbol{\eta}_2 \right\} f_{\eta_1|X}\left(\boldsymbol{\eta}_1 \mid \boldsymbol{X}\right) \mathrm{d}\boldsymbol{\eta}_1 = 0$$

for all $\boldsymbol{X}$ and $\boldsymbol{Y}$. Because $\boldsymbol{\eta}_1$ is complete sufficient in $F_{\eta_1|X}$,

$$\int f_{Y|\eta}\left(\boldsymbol{Y} \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\right) f_{\eta_2|\eta_1}\left(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1\right) \mathrm{d}\boldsymbol{\eta}_2 = \int \tilde{f}_{Y|\eta}\left(\boldsymbol{Y} \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\right) \tilde{f}_{\eta_2|\eta_1}\left(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1\right) \mathrm{d}\boldsymbol{\eta}_2 \quad \forall \boldsymbol{Y}, \boldsymbol{\eta}_1.$$

By assumption, $(f_{Y|\eta}, f_{\eta_2|\eta_1}) = (\tilde{f}_{Y|\eta}, \tilde{f}_{\eta_2|\eta_1})$. Therefore, Model A is identifiable.

To show the complete sufficiency of $\boldsymbol{\eta}_1$ under the sufficient condition, note that the density of $\boldsymbol{\eta}_1 \mid \boldsymbol{X}$ is of the form

$$f_{\eta_1|X}\left(\boldsymbol{\eta}_1 \mid \boldsymbol{X}\right) \quad \propto \quad f_{X|\eta_1}\left(\boldsymbol{X} \mid \boldsymbol{\eta}_1\right) f_{\eta_1}\left(\boldsymbol{\eta}_1\right) \propto \exp\left\{ \sum_{j=1}^{q} X_j s_j\left(\boldsymbol{\eta}_1\right) \right\} f^*\left(\boldsymbol{\eta}_1\right),$$

where f^* is a function of $\boldsymbol{\eta}_1$ that does not involve $\boldsymbol{X}$. Thus, as a property of the exponential family, $s(\boldsymbol{\eta}_1) \equiv (s_1(\boldsymbol{\eta}_1), \ldots, s_q(\boldsymbol{\eta}_1))$ is complete sufficient under the model with parameter $\boldsymbol{X} \in \mathcal{X}$. Because s is a one-to-one function, $\boldsymbol{\eta}_1$ is complete sufficient. □

Proof of Lemma 2.2. With an abuse of notation, we use $\boldsymbol{\nu}$ to denote all parameters in F_Y and F_η and drop the parameter vector in the arguments of the density functions. We consider the one-dimensional submodel along $(\boldsymbol{h}_\vartheta, \boldsymbol{h}_\beta, \boldsymbol{h}_\alpha, \boldsymbol{h}_\phi, \boldsymbol{h}_\nu, h_1(\cdot), \ldots, h_K(\cdot))$ for parameters $(\boldsymbol{\vartheta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\nu}, \Lambda_1, \ldots, \Lambda_K)$, where $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K)$, and $\boldsymbol{h}_\vartheta \equiv (\boldsymbol{h}_{\vartheta 1}, \ldots, \boldsymbol{h}_{\vartheta K})$ is partitioned accordingly. We define $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\phi}$, $\boldsymbol{h}_\beta$, $\boldsymbol{h}_\alpha$, and $\boldsymbol{h}_\phi$ in the same way. A one-dimensional submodel along the direction $h \equiv (\boldsymbol{h}_\theta, h_1(\cdot), \ldots, h_K(\cdot)) \in \mathbb{R}^d \times BV[0,\tau]^K$ indexed by ϵ is constructed by setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \epsilon \boldsymbol{h}_\theta$ for the vector of all Euclidean parameters $\boldsymbol{\theta}$ and $\Lambda_k(\cdot) = \int_0^{(\cdot)} \{1 + \epsilon h_k(s)\} \mathrm{d}\Lambda_k(s)$ for $k = 1, \ldots, K$. By the arguments in the proof of Theorem 2.1, we can consider the likelihood with the survival times being right censored at any values within $[0, \tau]$. For an observation with the K survival times right censored at $(t_1, \ldots, t_K)$, the likelihood is given by the left-hand side of (2.5). For simplicity of description, assume that m_k is the Lebesgue measure. If the score is zero almost surely, then

$$\int \int \exp\left\{-H\left(\boldsymbol{t}\right)\right\} g\left(\boldsymbol{s}\right) \frac{\partial}{\partial \boldsymbol{\nu}} \left\{ f_Y\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\eta}\right) f_\eta\left(\boldsymbol{\eta} \mid \boldsymbol{Z}\right) \right\}^{\mathrm{T}} \boldsymbol{h}_\nu \, \mathrm{d}\boldsymbol{s} \, \mathrm{d}\boldsymbol{\eta}$$

$$-\sum_{k=1}^{K}\int\int \exp\{-H(\boldsymbol{t})\}\, g(\boldsymbol{s})\, f_Y(\boldsymbol{Y}\mid \boldsymbol{Z},\boldsymbol{\eta})\, f_\eta(\boldsymbol{\eta}\mid \boldsymbol{Z})\, s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}$$
$$\times\int_0^{t_k} \boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}_{\vartheta k}+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{h}_{\phi k}+h_k(\omega)\,\mathrm{d}\Lambda_k(\omega)\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}\boldsymbol{\eta}=0 \quad (2.13)$$

for all $t_1,\ldots,t_K$, $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$, where $H(\boldsymbol{t})=\sum_{k=1}^{K}\Lambda_k(t_k)s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}$, $\boldsymbol{s}=(s_1,\ldots,s_K)^{\mathrm{T}}$, and $g(\boldsymbol{s})=\prod_{k=1}^{K}g_k(s_k)$. For $k=1,\ldots,K_1$, we differentiate (2.13) with respect to t_k and then set $t_l\to 0$ for $l=1,\ldots,K$. Thus,

$$\int\int g(\boldsymbol{s})\frac{\partial}{\partial\boldsymbol{\nu}}\{f_Y(\boldsymbol{Y}\mid\boldsymbol{Z},\boldsymbol{\eta})\, f_\eta(\boldsymbol{\eta}\mid\boldsymbol{Z})\}^{\mathrm{T}}\boldsymbol{h}_\nu s_k e^{\eta_{1k}}\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}\boldsymbol{s}$$
$$-\int\int s_k e^{\eta_{1k}} g(\boldsymbol{s})\, f_Y(\boldsymbol{Y}\mid\boldsymbol{Z},\boldsymbol{\eta})\, f_\eta(\boldsymbol{\eta}\mid\boldsymbol{Z})\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}\boldsymbol{s}\left\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}_{\vartheta k}+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+h_k(0)\right\}=0.$$

By linear independence of $\boldsymbol{W}$, $\boldsymbol{h}_{\vartheta k}=\boldsymbol{0}$.

Consider the first survival time T_1. Because $e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_1}$ takes at least two distinct values by conditions (C1) and (C3), we assume, without loss of generality, that it takes 1 and c with $c<1$. Let $U_1=s_1e^{\eta_{11}}$, and let f_{Y,U_1} be the density of $(\boldsymbol{Y},U_1)$ given $\boldsymbol{Z}$. Setting $t_2,\ldots,t_K\to 0$ and $e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_1}=1$ in (2.13), we have

$$\int_0^{t_1} h_1(\omega)\,\mathrm{d}\Lambda_1(\omega)=\left[\int\exp\left\{-\Lambda_1(t_1)U_1e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}\right\}U_1e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}f_{Y,U_1}(\boldsymbol{Y},U_1\mid\boldsymbol{Z})\,\mathrm{d}U_1\right]^{-1}$$
$$\int\exp\left\{-\Lambda_1(t_1)U_1e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}\right\}\left\{\frac{\partial}{\partial\boldsymbol{\nu}}f_{Y,U_1}(\boldsymbol{Y},U_1\mid\boldsymbol{Z})^{\mathrm{T}}\boldsymbol{h}_\nu\right.$$
$$\left.-\Lambda_1(t_1)\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta1}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha1}\right)U_1e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}f_{Y,U_1}(\boldsymbol{Y},U_1\mid\boldsymbol{Z})\right\}\mathrm{d}U_1$$
$$\equiv a\{\Lambda_1(t_1)\}. \quad (2.14)$$

Likewise, setting $t_2,\ldots,t_K\to 0$ and $e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_1}=c$ in (2.13), we have $ca\{\Lambda_1(t_1)\}=a\{c\Lambda_1(t_1)\}$. Thus, for all $v\in[0,\Lambda_1(\tau)]$, $a'(c^nv)=a'(v)$ for any integer n. It follows that $a'(v)=a'(0)$, such that a is a linear function. Let $a(v)=\kappa_1 v$. Then, with $v=\Lambda_1(t_1)$, (2.14) becomes

$$\kappa_1 v\int e^{-vU_1e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}}U_1e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}f_{Y,U_1}(\boldsymbol{Y},U_1\mid\boldsymbol{Z})\,\mathrm{d}U_1$$

$$= \int e^{-vU_1 e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}} \Big\{ \frac{\partial}{\partial \boldsymbol{\nu}} f_{Y,U_1}\left(\boldsymbol{Y}, U_1 \mid \boldsymbol{Z}\right)^{\mathrm{T}} \boldsymbol{h}_\nu$$
$$-v\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta 1}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha 1}\right) U_1 e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1} f_{Y,U_1}\left(\boldsymbol{Y}, U_1 \mid \boldsymbol{Z}\right) \Big\} \mathrm{d}U_1.$$

Therefore,

$$\int e^{-vU_1 e^{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_1+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_1}} \Big[\frac{\partial}{\partial \boldsymbol{\nu}} f_{Y,U_1}\left(\boldsymbol{Y}, U_1 \mid \boldsymbol{Z}\right)^{\mathrm{T}} \boldsymbol{h}_\nu$$
$$-\left(\kappa_1+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta 1}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha 1}\right) \frac{\partial}{\partial U_1}\left\{U_1 f_{Y,U_1}\left(\boldsymbol{Y}, U_1 \mid \boldsymbol{Z}\right)\right\} \Big] \mathrm{d}U_1 = 0$$

for all $v \in [0, \Lambda_1(\tau)]$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$. By the uniqueness of the Laplace transform, the term in the square brackets in the above integral is zero for all U_1, $\boldsymbol{Z}$, and $\boldsymbol{Y}$. Let $f_{Y,\bar{\eta}_1}$ be the density of $(\boldsymbol{Y}, \eta_{11})$ given $\boldsymbol{Z}$. Then,

$$\int \Big\{ \frac{1}{U_1} \frac{\partial}{\partial \boldsymbol{\nu}} f_{Y,\bar{\eta}_1}\left(\boldsymbol{Y}, \log U_1 - \log s \mid \boldsymbol{Z}\right)^{\mathrm{T}} \boldsymbol{h}_\nu$$
$$-\left(\kappa_1+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta 1}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha 1}\right) \frac{\partial}{\partial U_1} f_{Y,\bar{\eta}_1}\left(\boldsymbol{Y}, \log U_1 - \log s \mid \boldsymbol{Z}\right) \Big\} g_1(s)\, \mathrm{d}s = 0.$$

By the arguments for convolution in the proof of Theorem 2.1,

$$\frac{\partial}{\partial \boldsymbol{\nu}} f_{Y,\bar{\eta}_1}\left(\boldsymbol{Y}, \eta_{11} \mid \boldsymbol{Z}\right)^{\mathrm{T}} \boldsymbol{h}_\nu - \left(\kappa_1+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta 1}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha 1}\right) \frac{\partial}{\partial \eta_{11}} f_{Y,\bar{\eta}_1}\left(\boldsymbol{Y}, \eta_{11} \mid \boldsymbol{Z}\right) = 0$$

for all η_{11}. Multiplying both sides of the above equation by η_{11} and then integrating with respect to $(\boldsymbol{Y}, \eta_{11})$ at $\boldsymbol{Z} = \boldsymbol{0}$, we obtain

$$\frac{\partial}{\partial \boldsymbol{\nu}} \mathrm{E}\left(\eta_{11} \mid \boldsymbol{Z} = \boldsymbol{0}\right)^{\mathrm{T}} \boldsymbol{h}_\nu + \mathrm{E}\left(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha 1} \mid \boldsymbol{Z} = \boldsymbol{0}\right) + \kappa_1 = 0.$$

It then follows from condition (C2) that $\kappa_1 = 0$. Thus, $h_1(\cdot) = 0$.

Assume that $h_{k-1}(\cdot)$ has been shown to be a zero function for some $k = 2, \ldots, K_1$. Let $\bar{\boldsymbol{\eta}}_k = (\eta_{11}, \ldots, \eta_{1k})$, and let $f_{Y,\bar{\eta}_k}$ be the density of $(\boldsymbol{Y}, \bar{\boldsymbol{\eta}}_k)$ given $\boldsymbol{Z}$. Setting $t_{k+1}, \ldots, t_K \to 0$ in (2.13), we have

$$\int\int \exp\left\{-\sum_{l=1}^{k}\Lambda_l(t_l)s_l e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_l+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_l+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_l+\eta_{1l}}\right\} g(\boldsymbol{s})\left[\frac{\partial}{\partial\boldsymbol{\nu}}\left\{f_{Y,\bar{\eta}_k}(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k\mid\boldsymbol{Z})\right\}^{\mathrm{T}}\boldsymbol{h}_\nu\right.$$
$$-\sum_{l=1}^{k-1}\Lambda_l(t_l)\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta l}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha l}\right)s_l e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_l+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_l+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_l+\eta_{1l}}f_{Y,\bar{\eta}_k}(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k\mid\boldsymbol{Z})$$
$$\left.-s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\eta_{1k}}f_{Y,\bar{\eta}_k}(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k\mid\boldsymbol{Z})\int_0^{t_k}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+h_k(\omega)\,\mathrm{d}\Lambda_k(\omega)\right]\mathrm{d}\bar{\boldsymbol{\eta}}_k\,\mathrm{d}\boldsymbol{s}$$

equals zero for all $t_1,\ldots,t_k$, $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$. By the uniqueness of the Laplace transform, we conclude that $h_k(\omega)=\kappa_k\omega$ for some $\kappa_k\in\mathbb{R}$ and

$$\frac{\partial}{\partial\boldsymbol{\nu}}f_{Y,\bar{\eta}_k}(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k\mid\boldsymbol{Z})^{\mathrm{T}}\boldsymbol{h}_\nu+\kappa_k\frac{\partial}{\partial\eta_{1k}}f_{Y,\bar{\eta}_k}(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k\mid\boldsymbol{Z})$$
$$-\sum_{l=1}^{k}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta l}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha l}\right)\frac{\partial}{\partial\eta_{1l}}f_{Y,\bar{\eta}_k}(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k\mid\boldsymbol{Z})=0.$$

Multiplying both sides of the above equation by $(\eta_{11}\times\cdots\times\eta_{1k})$ and then integrating with respect to $(\boldsymbol{Y},\bar{\boldsymbol{\eta}}_k)$ at $\boldsymbol{Z}=\boldsymbol{0}$, we have $\kappa_k=0$ and $h_k(\cdot)=0$. By induction,

$$\frac{\partial}{\partial\boldsymbol{\nu}}f_{Y,\eta_1}(\boldsymbol{Y},\boldsymbol{\eta}_1\mid\boldsymbol{Z})^{\mathrm{T}}\boldsymbol{h}_\nu-\sum_{k=1}^{K_1}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}\right)\frac{\partial}{\partial\eta_{1k}}f_{Y,\eta_1}(\boldsymbol{Y},\boldsymbol{\eta}_1\mid\boldsymbol{Z})=0\quad\forall\boldsymbol{Z},\boldsymbol{Y},\boldsymbol{\eta}_1.$$

It then follows from condition (D5) that $\boldsymbol{h}_\nu=\boldsymbol{0}$, $\boldsymbol{h}_{\beta k}=\boldsymbol{0}$, and $\boldsymbol{h}_{\alpha k}=\boldsymbol{0}$ for $k=1,\ldots,K_1$.

Consider the left-hand side of (2.13). The first term and the first K_1 terms in the summation of the second term have been shown to be zero. Thus, the left-hand side of (2.13) can be viewed as Laplace transforms with arguments $\Lambda_1(t_1),\ldots,\Lambda_{K_1}(t_{K_1})$. By the properties of the Laplace transform and function convolution,

$$-\sum_{k=K_1+1}^{K}\int\int\exp\left\{-\sum_{l=K_1+1}^{K}\Lambda_l(t_l)\,s_l e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_l+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_l+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_l+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_l}\right\}$$
$$\times\left\{\prod_{l=K_1+1}^{K}g_l(s_l)\right\}f_Y(\boldsymbol{Y}\mid\boldsymbol{Z},\boldsymbol{\eta})\,f_\eta(\boldsymbol{\eta}\mid\boldsymbol{Z})\,s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\,\mathrm{d}(s_{K_1+1},\ldots,s_K)$$
$$\times\left\{\int_0^{t_k}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}_{\vartheta k}+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{h}_{\phi k}+h_k(\omega)\,\mathrm{d}\Lambda_k(\omega)\right\}\mathrm{d}\boldsymbol{\eta}_2=0$$

for all $t_{K_1+1},\ldots,t_K$, $\boldsymbol{\eta}_1$, $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$. For $k=K_1+1,\ldots,K$, setting $t_l\to0$ for $l\neq k$ in the

above equation yields

$$\int\int e^{-\Lambda_k(t_k)s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}} s_k e^{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\vartheta}_k+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_k+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} f_{\eta_2^{(k)}|Y,\eta_1}(\boldsymbol{\eta}_2^{(k)} \mid \boldsymbol{Z},\boldsymbol{Y},\boldsymbol{\eta}_1)$$
$$\times g_k(s_k)\left\{\int_0^{t_k} \boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}_{\vartheta k}+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{h}_{\phi k}+h_k(\omega)\,\mathrm{d}\Lambda_k(\omega)\right\}\mathrm{d}\boldsymbol{\eta}_2^{(k)}\,\mathrm{d}s_k=0$$

for all t_k, $\boldsymbol{\eta}_1$, $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$, where $f_{\eta_2^{(k)}|Y,\eta_1}$ is the density of $\boldsymbol{\eta}_2^{(k)}$ given $(\boldsymbol{Z},\boldsymbol{Y},\boldsymbol{\eta}_1)$. By condition (C5), $\boldsymbol{\eta}_2^{(k)}$ is complete sufficient in $f_{\eta_2^{(k)}|Y,\eta_1}$ with $(\boldsymbol{Y}^{-(k)},\boldsymbol{\eta}^{-(k)})$ as parameters. Because $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\alpha}_k+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k$ and $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{h}_{\phi k}$ do not depend on $(\boldsymbol{Y}^{-(k)},\boldsymbol{\eta}^{-(k)})$, the property of complete sufficient statistics implies that

$$\int_0^{t_k} \boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}_{\vartheta k}+\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{h}_{\beta k}+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{h}_{\alpha k}+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{h}_{\phi k}+h_k(\omega)\,\mathrm{d}\Lambda_k(\omega)=0 \quad \forall t_k\in[0,\tau],\boldsymbol{\eta},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{Y}.$$

Thus, $\boldsymbol{h}_{\vartheta k}=\mathbf{0}$, $\boldsymbol{h}_{\beta k}=\mathbf{0}$, $\boldsymbol{h}_{\alpha k}=\mathbf{0}$, and $\boldsymbol{h}_{\phi k}=\mathbf{0}$. Therefore, $h_k(\cdot)=0$.

We have shown that the information operator is one-to-one. Using the arguments of Zeng and Lin (2010), we can show that it is also a Fredholm operator. Therefore, the invertibility of the information operator follows. □

Proof of Lemma 2.3. Without loss of generality, assume that Y is a scalar. Clearly,

$$h(Y,\boldsymbol{b})\,g(Y,\boldsymbol{b}) \;\le\; \prod_{j=1}^{q}\exp\{K(1+|Y|)+(M_j+K)|b_j|+N_jYb_j-c_jb_j^2\}.$$

Note that $(M+K)|b_j|+NYb_j-cb_j^2$ is bounded by

$$-c\left(\left|b_j-\frac{NY}{2c}\right|-\frac{M+K}{2c}\right)^2+\frac{(M+K)^2}{4c}+(M+K)\left|\frac{NY}{2c}\right|+\frac{N^2Y^2}{4c}.$$

Similarly,

$$-M|b_j|+NYb_j-cb_j^2 \;\ge\; -c\left(\left|b_j-\frac{NY}{2c}\right|+\frac{M}{2c}\right)^2-M\left|\frac{NY}{2c}\right|+\frac{N^2Y^2}{4c}.$$

Therefore, (2.12) is bounded by

$$e^{O(1+|Y|)}\frac{\int\prod_{j=1}^q e^{-c_j\left(\left|b_j-\frac{N_jY}{2c_j}\right|-\frac{M_j+K}{2c_j}\right)^2}\,\mathrm{d}\boldsymbol{b}}{\int\prod_{j=1}^q e^{-c_j\left(\left|b_j-\frac{N_jY}{2c_j}\right|+\frac{M_j}{2c_j}\right)^2}\,\mathrm{d}\boldsymbol{b}}=e^{O(1+|Y|)}\frac{\int\prod_{j=1}^q e^{-c_j\left(|b_j|-\frac{M_j+K}{2c_j}\right)^2}\,\mathrm{d}\boldsymbol{b}}{\int\prod_{j=1}^q e^{-c_j\left(|b_j|+\frac{M_j}{2c_j}\right)^2}\,\mathrm{d}\boldsymbol{b}}\le e^{O(1+|Y|)}.$$

Likewise, (2.11) is bounded by

$$e^{O(1+|Y|)}\frac{\int\exp\left\{-e^{A_1+\left(B+\sum_j N_j/2c_j\right)Y+\mathbf{1}_q^{\mathrm{T}}\boldsymbol{b}}\right\}\prod_{j=1}^q e^{-c_j\left(|b_j|+\frac{M_j+K}{2c_j}\right)^2}\,\mathrm{d}\boldsymbol{b}}{\int\exp\left\{-e^{A_2+\left(B+\sum_j N_j/2c_j\right)Y+\mathbf{1}_q^{\mathrm{T}}\boldsymbol{b}}\right\}\prod_{j=1}^q e^{-c_j\left(|b_j|+\frac{M_j}{2c_j}\right)^2}\,\mathrm{d}\boldsymbol{b}}. \tag{2.15}$$

For any $w>0$ and $a\in\mathbb{R}$,

$$\int_{b\in\mathbb{R}}\exp\left(-we^b\right)e^{-(|b|-a)^2}\,\mathrm{d}b \;\le\; 2\int_{b\in\mathbb{R}}\exp\left(-we^{-a}e^{-b}\right)e^{-b^2}\,\mathrm{d}b.$$

Thus, the numerator in (2.15) is bounded above by

$$2^q\int\exp\left\{-e^{A_1+\sum_j -(M_j+K)/2c_j+(B+N_j/2c_j)Y-\mathbf{1}_q^{\mathrm{T}}\boldsymbol{b}}\right\}\prod_{j=1}^q e^{-c_jb_j^2}\,\mathrm{d}\boldsymbol{b}.$$

In addition, if $a>0$, then

$$\int_{b\in\mathbb{R}}\exp\left(-we^b\right)e^{-(|b|+a)^2}\,\mathrm{d}b \;\ge\; \frac{1}{2}K_a^{-1}\int_{b\in\mathbb{R}}\exp\left(-we^ae^{-b}\right)e^{-b^2}\,\mathrm{d}b,$$

where $K_a=\int_0^\infty e^{-b^2}\,\mathrm{d}b/\int_a^\infty e^{-b^2}\,\mathrm{d}b$. Thus, the denominator in (2.15) is bounded below by

$$\frac{1}{2^q}\prod_{j=1}^q K_{M_j/2c_j}^{-1}\int\exp\left\{-e^{A_2+\sum_j M_j/2c_j+(B+N_j/2c_j)Y-\mathbf{1}_q^{\mathrm{T}}\boldsymbol{b}}\right\}\prod_{j=1}^q e^{-c_jb_j^2}\,\mathrm{d}\boldsymbol{b}.$$

The fraction in (2.15) is bounded by

$$4^q\prod_{j=1}^q K_{M_j/2c_j}\frac{\int\exp\left\{-w_1e^{-\left(\boldsymbol{c}^{-1/2}\right)^{\mathrm{T}}\boldsymbol{b}}\right\}e^{-|\boldsymbol{b}|^2}\,\mathrm{d}\boldsymbol{b}}{\int\exp\left\{-w_2e^{-\left(\boldsymbol{c}^{-1/2}\right)^{\mathrm{T}}\boldsymbol{b}}\right\}e^{-|\boldsymbol{b}|^2}\,\mathrm{d}\boldsymbol{b}},$$

where $\boldsymbol{c}^{-1/2}=(c_1^{-1/2},\ldots,c_q^{-1/2})^{\mathrm{T}}$, and $w_k=e^{A_k+\sum_j(-1)^k(M_j+K)/2c_j+(B+N_j/2c_j)Y}$ for $k=1,2$.

Therefore, the fraction in (2.15) is finite if $\sum_j (B + N_j/2c_j)Y \to -\infty$. If $\sum_j (B + N_j/2c_j)Y \to \infty$, then we use the approximation that

$$\begin{aligned}
& \int_{b\in\mathbb{R}} \exp\left(-\frac{b^2}{2} - we^{\mu b}\right) \mathrm{d}b \\
= \; & \left[\frac{2\pi\left\{1+o\left(1\right)\right\}}{\log w}\right]^{1/2} \exp\Bigg(-\frac{1}{2\mu^2}\left[\log w\left\{1 - \frac{\log\log w - \log\mu^2}{\log w} + o\left(\frac{1}{\log w}\right)\right\}\right]^2 \\
& -\frac{\log w}{\mu^2}\left\{1 - \frac{\log\log w - \log\mu^2}{\log w} + o\left(\frac{1}{\log w}\right)\right\}\Bigg)
\end{aligned}$$

as $w \to \infty$ (Evans and Swartz 2000). It follows that

$$\begin{aligned}
& \frac{\int \exp\left\{-w_1 e^{-\left(\boldsymbol{c}^{-1/2}\right)^{\mathrm{T}}\boldsymbol{b}}\right\} e^{-|\boldsymbol{b}|^2}\,\mathrm{d}\boldsymbol{b}}{\int \exp\left\{-w_2 e^{-\left(\boldsymbol{c}^{-1/2}\right)^{\mathrm{T}}\boldsymbol{b}}\right\} e^{-|\boldsymbol{b}|^2}\,\mathrm{d}\boldsymbol{b}} \\
= \; & \frac{\int \exp\left(-w_1 e^{-|\boldsymbol{c}^{-1/2}|b}\right) e^{-b^2}\,\mathrm{d}b}{\int \exp\left(-w_2 e^{-|\boldsymbol{c}^{-1/2}|b}\right) e^{-b^2}\,\mathrm{d}b} \\
= \; & O\left(1\right)\left(\frac{\log w_2}{\log w_1}\right)^{1/2} \exp\Bigg[-\frac{1}{2\left|\boldsymbol{c}^{-1/2}\right|^2}\left\{\log w_1 - \log\log w_1 + 2\log\left|\boldsymbol{c}^{-1/2}\right| + o\left(1\right)\right\}^2 \\
& -\frac{1}{|\boldsymbol{c}^{-1}|^2}\left\{\log w_1 - \log\log w_1 + 2\log\left|\boldsymbol{c}^{-1/2}\right| + o\left(1\right)\right\} \\
& +\frac{1}{2\left|\boldsymbol{c}^{-1/2}\right|^2}\left\{\log w_2 - \log\log w_2 + 2\log\left|\boldsymbol{c}^{-1/2}\right| + o\left(1\right)\right\}^2 \\
& +\frac{1}{|\boldsymbol{c}^{-1}|^2}\left\{\log w_2 - \log\log w_2 + 2\log\left|\boldsymbol{c}^{-1/2}\right| + o\left(1\right)\right\}\Bigg].
\end{aligned}$$

The Y^2 terms in the exponent cancel out; therefore, (2.11) is bounded by $e^{O(1+|Y|)}$. □

Proof of Lemma 2.4. We use $\boldsymbol{Z}$ to denote both $\boldsymbol{W}$ and $\boldsymbol{Z}$ with $\boldsymbol{\beta}_k$ $(k = 1, \ldots, K)$ as the corresponding vector of regression parameters. Note that

$$\begin{aligned}
& \frac{\partial}{\partial\boldsymbol{\beta}} \log\Psi(\mathcal{O}_i;\boldsymbol{\theta},\mathcal{A}) \qquad (2.16) \\
= \; & \int \Omega_{ki}\left(\boldsymbol{\eta}\right)\Bigg(\Delta_{ki}\Big[1 + \frac{G_k''\left\{q_{ki}\left(\boldsymbol{\theta},\mathcal{A}\right)\right\}}{G_k'\left\{q_{ki}\left(\boldsymbol{\theta},\mathcal{A}\right)\right\}}\Lambda_k(\tilde{T}_{ki})e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\Big] \\
& -G_k'\left\{q_{ki}\left(\boldsymbol{\theta},\mathcal{A}\right)\right\}\Lambda_k(\tilde{T}_{ki})e^{\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\Bigg)\boldsymbol{Z}_i f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i,\boldsymbol{\eta};\boldsymbol{\psi}) f_\eta(\boldsymbol{\eta} \mid \boldsymbol{Z}_i;\boldsymbol{\nu})\,\mathrm{d}\boldsymbol{\eta}
\end{aligned}$$

$$\times \left\{ \int \Omega_{ki}(\boldsymbol{\eta}) f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\eta}; \boldsymbol{\psi}) f_\eta(\boldsymbol{\eta} \mid \boldsymbol{Z}_i; \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\eta} \right\}^{-1},$$

where $q_{ki}(\boldsymbol{\theta}, \mathcal{A}) = e^{\boldsymbol{Z}_i^{\mathrm{T}}\beta_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k}\Lambda_k(\tilde{T}_{ki})$, and

$$\Omega_{ki}(\boldsymbol{\eta}) = \left[e^{\boldsymbol{Z}_i^{\mathrm{T}}\beta_k + \boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{\alpha}_k + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\phi}_k} G_k'\{q_{ki}(\boldsymbol{\theta}, \mathcal{A})\} \right]^{\Delta_{ki}} \exp\left[-G_k\{q_{ki}(\boldsymbol{\theta}, \mathcal{A})\} \right].$$

By condition (D4), the terms $G_k'\{q_{ki}(\boldsymbol{\theta}, \mathcal{A})\}$ and $[G_k'\{q_{ki}(\boldsymbol{\theta}, \mathcal{A})\}]^{-1} G_k''\{q_{ki}(\boldsymbol{\theta}, \mathcal{A})\}$ are bounded by $e^{O(1+|\boldsymbol{Y}_i|+|\boldsymbol{\eta}|)}$. Therefore, the first integral term on the right-hand side of (2.16) is bounded above by

$$\int \Omega_{ki}(\boldsymbol{\eta}) \, e^{O(1+|\boldsymbol{Y}_i|+|\boldsymbol{\eta}|)} f_Y(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{\eta}; \boldsymbol{\psi}) f_\eta(\boldsymbol{\eta} \mid \boldsymbol{Z}_i; \boldsymbol{\nu}) \, \mathrm{d}\boldsymbol{\eta}.$$

Condition (D3) states that either $G_k(x)/\log x \to M_k$ or $G_k(x)/x^{\rho_k} \to M_k$ as $x \to \infty$. If $G_k(x)/x^{\rho_k} \to M_k$, then we can find $M_{1k} > 0$, $M_{2k} > 0$, $C_{1k} \in \mathbb{R}$, and $C_{2k} \in \mathbb{R}$ such that $M_{2k}x^{\rho_k} + C_{2k} \le G_k(x) \le M_{1k}x^{\rho_k} + C_{1k}$. Thus,

$$\exp\left(-e^{A_1 + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}_i + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\eta}} \right) e^{-O(1+|\boldsymbol{Y}_i|+|\boldsymbol{\eta}|)} \le \Omega_{ki}(\boldsymbol{\eta}) \le \exp\left(-e^{A_2 + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}_i + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\eta}} \right) e^{O(1+|\boldsymbol{Y}_i|+|\boldsymbol{\eta}|)}$$

for some A_1, A_2, $\boldsymbol{B}$, and $\boldsymbol{C}$. After transforming $\boldsymbol{\eta}$ to $\boldsymbol{b}$ using the transformation S specified in condition (D3), we see that (2.16) is bounded by a term of the form (2.11) and is in turn bounded by $e^{O(1+|\boldsymbol{Y}_i|)}$ according to Lemma 2.3. If $G_k(x)/\log x \to M_k$, then (2.16) is bounded by a term of the form (2.12), which is also bounded by $e^{O(1+|\boldsymbol{Y}_i|)}$. Similarly, the derivatives of $\log \Psi(\mathcal{O}_i; \boldsymbol{\theta}, \mathcal{A})$ with respect to other parameters are bounded by $e^{O(1+|\boldsymbol{Y}_i|)}$.

By Theorem 2.7.5 of van der Vaart and Wellner (1996), the bracket covering number for any bounded set in $BV[0, \tau]$ is of the order $\exp\{O(1/\epsilon)\}$. Therefore, we can construct $N_\epsilon \equiv 1/\epsilon^d \times \exp\{O(K/\epsilon)\}$ brackets for $\Theta \times BV[0, \tau]^K$, denoted by $\{(\boldsymbol{\theta}_j^L, \mathcal{A}_j^L), (\boldsymbol{\theta}_j^U, \mathcal{A}_j^U)\}$ $(j = 1, \ldots, N_\epsilon)$, where $|\boldsymbol{\theta}_j^U - \boldsymbol{\theta}_j^L| < \epsilon$, and

$$\int_0^\tau \left| \Lambda_{kj}^U(t) - \Lambda_{kj}^L(t) \right|^2 \mathrm{E}\left\{ e^{O(1+|\boldsymbol{Y}_i|)} \, \mathrm{d}I(\tilde{T}_{ki} \le t) \right\} < \epsilon^2.$$

By the mean-value theorem, for any $(\boldsymbol{\theta}_1, \mathcal{A}_1)$ and $(\boldsymbol{\theta}_2, \mathcal{A}_2)$,

$$\left|\log \Psi(\mathcal{O}_i; \boldsymbol{\theta}_1, \mathcal{A}_1) - \log \Psi(\mathcal{O}_i; \boldsymbol{\theta}_2, \mathcal{A}_2)\right| \leq e^{O(1+|\boldsymbol{Y}_i|)} \left\{ |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + \sum_{k=1}^{K} \left| \Lambda_{1k}(\tilde{T}_{ki}) - \Lambda_{2k}(\tilde{T}_{ki}) \right| \right\}.$$

The pairs of functions

$$\log \Psi(\mathcal{O}_i; \boldsymbol{\theta}_j^L, \mathcal{A}_j^L) \pm e^{O(1+|\boldsymbol{Y}_i|)} \left\{ \left| \boldsymbol{\theta}_j^U - \boldsymbol{\theta}_j^L \right| + \sum_{k=1}^{K} \left| \Lambda_{kj}^U(\tilde{T}_{ki}) - \Lambda_{kj}^L(\tilde{T}_{ki}) \right| \right\},$$

$j = 1, \ldots, N_\epsilon$, constitute a bracket cover for $\mathcal{C}_1$, where the $L_2(\mathcal{P})$-distance within each bracket pair is of the order ϵ. Therefore, the bracket entropy of $\mathcal{C}_1$ is finite, such that $\mathcal{C}_1$ is Donsker. Similarly, the classes $\mathcal{C}_2$ and $\mathcal{C}_{3k}$ can also be shown to be Donsker. □

CHAPTER 3

ROBUST SCORE TESTS WITH MISSING DATA IN MULTI-PLATFORM GENOMICS STUDIES

3.1 Introduction

Recent technological advances have made it possible to measure multiple genomics platforms on the same set of subjects. However, constraints regarding cost and other factors prohibit measurement of all platforms on all study subjects. For example, in The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), over 10,000 subjects with 33 cancer types were measured on multiple genomics platforms, including somatic mutation, copy number variation, and expressions of microRNA, mRNA, and protein, but for a substantial number of subjects, data on RNA sequencing and protein expressions were not generated. As another example, in the National Heart, Lung, and Blood Institute's Exome Sequencing Project (https://esp.gs.washington.edu/), only 7,000 subjects with specific diseases or conditions were selected for whole-exome sequencing from the tens of thousands of total subjects with genotyping array data (Lin et al. 2013). Finally, in the Trans-Omics for Precision Medicine (TOPMed) program (https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed) and the Genome Sequencing Program (GSP) (http:// gsp-hg.org), whole-genome sequencing data will be available on hundreds of thousands of subjects, but other genomics platforms, such as RNA sequencing, methylation, and metabolites, will be available for only a few thousand subjects through ancillary studies of specific diseases.

It is desirable to infer missing data on one genomics platform using available data from other platforms. Indeed, this has become a routine practice with genotype data, where linkage disequilibrium allows one to impute, with great accuracy, sequencing data from genotyping array data (Li et al. 2010; Auer et al. 2012). A far greater challenge is to infer missing values for a quantitative measurement, such as the expression of RNA or protein, from other quantitative measurements or from SNP genotype data due to the complex and noisy relationships among those variables (Kim et al. 2005; Torres-García et al. 2009).

Several authors have considered missing data in the context of association testing, which is of

primary interest in genomics studies. Specifically, Hu et al. (2015) studied the score test based on imputed genotype data and proposed a variance estimator that properly accounts for the differential quality between observed and imputed genotypes. The method requires that the imputation is unbiased and the genotype is independent of the other variables in the phenotype model. Derkach et al. (2015) and Lawless (2016) proposed to model the variable with missing values under outcome-dependent sampling and studied the score test based on the full likelihood. Derkach et al. (2015) assumed a nonparametric model for the variable with missing values and restricted covariates to only a few possible values. Lawless (2016) assumed a full parametric missing-data model. All existing methods require unbiased imputation or correct modeling of the variable with missing values. This is difficult to achieve, especially when the number of covariates in the missing-data model is not small.

In this chapter, we investigate the validity of the score test with imputed data when the missing-data mechanism may depend on the phenotype and covariates. In particular, we show that a condition weaker than correct specification of the missing-data model is sufficient for the score statistic to be unbiased. Based on this finding, we propose a robust score test which, unlike existing methods, preserves the type I error under general missing-data mechanisms even when the imputation model is misspecified. The proposed score statistic is based on a semiparametric model for the variable with missing values, where covariates enter the model linearly and also through a one-dimensional nonparametric function. As a result, the test is feasible with a large number of covariates in the missing-data model. The proposed methodology is applicable to all common phenotype models and encompasses continuous, binary, and right-censored phenotypes.

In Chapter 3.2, we formulate the problem, investigate the validity of the standard score test under various missing-data mechanisms, and develop the robust score test. In Chapter 3.3, we report results from simulation studies that compare the proposed and existing methods. In Chapter 3.4, we provide an application to a dataset from TCGA. We make concluding remarks in Chapter 3.5 and relegate technical details to the Chapter 3.6.

3.2 Methods

Consider a genomics study that involves phenotype Y, genomic variable of interest S, and vector of covariates $\boldsymbol{X}$. For example, Y may represent disease status, S may represent the RNA expression of a gene, and $\boldsymbol{X}$ may include genomic variables associated with S, such as the mutation status

and copy number of the gene, or non-genomic variables, such as tumor stage, age, and ancestry. Let $f(\cdot; \beta S + \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}, \boldsymbol{\zeta})$ denote the density function of Y conditional on $(S, \boldsymbol{X})$, where β and $\boldsymbol{\gamma}$ are regression parameters, and $\boldsymbol{\zeta}$ is a set of nuisance parameters that may be infinite-dimensional; this is referred to as the phenotype model. In particular, $\boldsymbol{\zeta}$ is the dispersion parameter in the generalized linear model and the baseline hazard function in the proportional hazards model. We allow S to be missing and let R indicate, by values of 1 versus 0, whether S is observed or not, respectively. Let $\boldsymbol{Z}$ be a set of predictors for S that includes $\boldsymbol{X}$, as well as variables that are not present in the phenotype model. The extra variables in $\boldsymbol{Z}$ are exogenous variables that affect Y indirectly through S and $\boldsymbol{X}$, such that $\boldsymbol{Z}$ is independent of Y conditional on $\boldsymbol{X}$ under $\beta = 0$. The observed data consist of $(Y_i, S_i R_i, R_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, n$.

We are interested in testing the null hypothesis $H_0 : \beta = 0$. Among the three common tests, namely the Wald's test, the likelihood ratio test, and the score test, the first two require fitting the model under the alternative hypothesis, which involves estimation of the conditional distribution of S given $\boldsymbol{X}$ in the presence of missing values for S. If the model for S is misspecified (which is inevitable when the dimension of $\boldsymbol{X}$ is moderately high), then the estimators of the nuisance parameters may be inconsistent, such that the resulting tests are invalid. By contrast, the score test only requires fitting the model under the null hypothesis. As a result, the score test requires fewer assumptions on the missing-data model than the other two tests in order to yield correct type I error. Therefore, we focus on the score test in the rest of this chapter.

The score statistic for β at $\beta = 0$ takes the form of $A(Y, \boldsymbol{X}; \boldsymbol{\psi})S$, where $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = \partial \log f(Y; t + \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}, \boldsymbol{\zeta}) / \partial t \mid_{t=0}$, and $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\zeta})$. Note that $\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0) \mid \boldsymbol{X}\} = 0$, where $\boldsymbol{\psi}_0 \equiv (\boldsymbol{\gamma}_0, \boldsymbol{\zeta}_0)$ is the true value of $\boldsymbol{\psi}$. This formulation includes many common models as special cases. For the linear model, $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = \sigma^{-2}(Y - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X})$, where σ^2 is the error variance. For the logistic model, $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = Y - e^{\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}} / (1 + e^{\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}})$. For the proportional hazards model with right censoring, $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = \Delta - \Lambda(\tilde{T}) e^{\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}}$, where $Y = (\tilde{T}, \Delta)$, $\tilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$, T is the survival time of interest, C is the censoring time, $I(\cdot)$ is the indicator function, and Λ is the cumulative baseline hazard function.

We consider the score statistic based on the imputed S. We specify an imputation model of S that depends on $\boldsymbol{Z}$ and a set of parameters $\boldsymbol{\xi}$. Let $\tilde{S}(\boldsymbol{Z}_i; \hat{\boldsymbol{\xi}})$ be the imputed value of S_i, where $\hat{\boldsymbol{\xi}}$ is

an estimator of $\boldsymbol{\xi}$. The (normalized) imputation "score" statistic is

$$U_\beta^{\text{imp}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = n^{-1/2} \sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\psi}}) \{R_i S_i + (1 - R_i) \tilde{S}(\boldsymbol{Z}_i; \hat{\boldsymbol{\xi}})\},$$

where $\hat{\boldsymbol{\psi}} \equiv (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\zeta}})$ is an estimator of $\boldsymbol{\psi}$ under H_0. At $\beta = 0$, the score statistic based on the full likelihood with a regression model of S on $\boldsymbol{Z}$ takes the form of U_β^{imp}. However, the proposed imputation score statistic is more general in that $\tilde{S}$ needs not be the posterior mean of S (given the observed data) evaluated at the maximum likelihood estimator of $\boldsymbol{\xi}$. Let $\boldsymbol{\xi}^*$ be the limit of $\hat{\boldsymbol{\xi}}$. The following proposition provides a general sufficient condition for the unbiasedness of the imputation score statistic under H_0.

Proposition 3.1. *Assume that there exists a subset of $\boldsymbol{X}$, denoted by $\tilde{\boldsymbol{X}}$, such that R is independent of $(S, \boldsymbol{Z})$ conditional on $(Y, \tilde{\boldsymbol{X}})$ and* $\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \tilde{\boldsymbol{X}}) = \mathrm{E}\{\tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \tilde{\boldsymbol{X}}\}$. *Then,* $\mathrm{E}\{U_\beta^{\text{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)\} = 0$ *under* $\beta = 0$.

The proofs of this proposition and other technical results are provided in Chapter 3.6.1.

Remark 3.1. The missing-data mechanism assumed in this proposition may arise from the extreme-tail sampling scheme, where only subjects with extreme values of Y are selected for measurements of S (Lin et al. 2013). In this case, the inverse probability weighting approach is not feasible because $P(R = 1 \mid Y)$ is zero for some subjects, whereas the imputation approach is applicable.

Remark 3.2. The dependence between R and $\tilde{\boldsymbol{X}}$ may be introduced in the design stage, when the sampling of S is performed separately at different values of $\tilde{\boldsymbol{X}}$. In cancer genomics, $\tilde{\boldsymbol{X}}$ may include risk factors such as tumor stage and tumor grade, and subjects with unusually high or unusually low risk may be more likely to be selected for measurements of S. In this case, $\tilde{\boldsymbol{X}}$ is a low-dimensional subset of $\boldsymbol{X}$ that is discrete, and a nonparametric modeling of S on $\tilde{\boldsymbol{X}}$ is feasible.

Remark 3.3. The condition in Proposition 3.1 requires that the true and imputed S variables have the same conditional expectation given $\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$. This condition is trivially satisfied if S is independent of $\boldsymbol{X}$ and the imputed value has the same mean as S, as assumed by Hu et al. (2015). For the score statistic to be unbiased, we only need the expectation of the true and imputed S variables conditional on $\tilde{\boldsymbol{X}}$ and a single index $\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}$ to be the same. This is practically achievable via nonparametric modeling of S given the low-dimensional covariates $(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \tilde{\boldsymbol{X}})$, even though the

whole set of covariates $\boldsymbol{X}$ may be high-dimensional. If the missing-data mechanism does not depend on covariates, then $\tilde{\boldsymbol{X}}$ is absent, such that it is only necessary to correctly model the conditional expectation of S given the single index $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$.

Proposition 3.1 implies that the imputation score statistic is unbiased under H_0 if the conditional expectation of S given a specific projection of $\boldsymbol{Z}$ (but not necessarily the full set of $\boldsymbol{Z}$) is correctly specified. To guarantee that this condition holds, we model the relationship between S and $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}})$ nonparametrically when $\tilde{\boldsymbol{X}}$ is discrete and takes a small number of values. Because the regression model of S on $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}})$ may not be very predictive, we include other components of $\boldsymbol{Z}$ in the imputation model in order to improve the imputation accuracy. Given the nonparametric function of $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}})$, the inclusion of $\boldsymbol{Z}$ will not result in bias of the score statistic even if the imputation model is misspecified. In the sequel, we assume that the missing-data mechanism specified in Proposition 3.1 holds and that $\tilde{\boldsymbol{X}}$ is discrete with possible values $(\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_L)$. For each $\tilde{\boldsymbol{x}}_l$ $(l = 1, \ldots, L)$, we assume the working model $\mathrm{E}(S \mid \boldsymbol{Z}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) = g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}$, where g_l is unspecified, and $\tilde{\boldsymbol{Z}}$ is a specific q-dimensional function of $\boldsymbol{Z}$ that is (asymptotically) orthogonal to $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}})$. Let $(g_l^*, \boldsymbol{\eta}_l^*) = \arg\min_{(g_l, \boldsymbol{\eta}_l)} \mathrm{E}[R\{S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}^2 | \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l]$ almost surely, $\boldsymbol{\xi} = (g_1, \ldots, g_L, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L)$, and $\boldsymbol{\xi}^* = (g_1^*, \ldots, g_L^*, \boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_L^*)$. The following proposition states the unbiasedness of the resulting imputation score statistic.

Proposition 3.2. *If* $\tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) = \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{*\mathrm{T}}\tilde{\boldsymbol{Z}}\}$, *then* $\mathrm{E}\{U_\beta^{\mathrm{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)\} = 0$.

Proposition 3.2 motivates us to estimate g_l and $\boldsymbol{\eta}_l$ using least-squares regression with the complete observations. We propose to approximate g_l $(l = 1, \ldots, L)$ with B-spline functions of order m (De Boor 1978) and replace the true value $\boldsymbol{\gamma}_0$ by the estimator $\hat{\boldsymbol{\gamma}}$. For simplicity of presentation, we assume the same set of fixed B-spline functions for each g_l, but we allow them to be chosen adaptively and separately for each g_l in practice. Let m and K_n be integers, such that $K_n \geq m \geq 2$, and K_n depends on the sample size n. For a set of grid points $\boldsymbol{\tau} \equiv (\tau_0, \ldots, \tau_{K_n - m + 1})$, such that $\min_{\boldsymbol{X}} \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X} = \tau_0 < \cdots < \tau_{K_n - m + 1} = \max_{\boldsymbol{X}} \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}$, let $\boldsymbol{B}(\cdot) = (B_1(\cdot), \ldots, B_{K_n}(\cdot))^{\mathrm{T}}$, where B_k is the kth m-order B-spline function on $\boldsymbol{\tau}$; the grid points at the two ends have multiplicity m. For $l = 1, \ldots, L$, let

$$(\hat{\boldsymbol{\alpha}}_l, \hat{\boldsymbol{\eta}}_l) = \underset{(\boldsymbol{\alpha}_l, \boldsymbol{\eta}_l)}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} R_i I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l) \Big\{ S_i - \sum_{k=1}^{K_n} \alpha_{lk} B_k(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}_i \Big\}^2,$$

where $\boldsymbol{\alpha}_l = (\alpha_{l1}, \ldots, \alpha_{lK_n})^{\mathrm{T}}$. Effectively, we partition the data into L strata, with each stratum corresponding to a value of $\tilde{\boldsymbol{x}}_l$, and we perform separate least-squares regression for each stratum using subjects with observed S. Let $\hat{\boldsymbol{\alpha}}_l = (\hat{\alpha}_{l1}, \ldots, \hat{\alpha}_{lK_n})^{\mathrm{T}}$, $\hat{g}_l = \sum_{k=1}^{K_n} \hat{\alpha}_{lk} B_k$, and $\hat{\boldsymbol{\xi}} = (\hat{g}_1, \ldots, \hat{g}_L, \hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_L)$. The robust imputation score statistic is $U_\beta^{\mathrm{rob}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}})$, where

$$U_\beta^{\mathrm{rob}}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma}) = n^{-1/2} \sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \boldsymbol{\psi}) \Big[R_i S_i + (1 - R_i) \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l) \{ g_l(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}_i) + \boldsymbol{\eta}_l^{\mathrm{T}} \tilde{\boldsymbol{Z}}_i \} \Big],$$

and the third argument in $U_\beta^{\mathrm{rob}}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ corresponds to $\boldsymbol{\gamma}$ in the argument of g_l.

Let $\ell_\beta(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ be $U_\beta^{\mathrm{rob}}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ for a single subject, $\ell_\psi(\boldsymbol{\psi})[\boldsymbol{h}_1]$ be the derivative of $\log f(Y; \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}, \boldsymbol{\zeta})$ along the path $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \epsilon \boldsymbol{h}_1$, with $\boldsymbol{h}_1$ being a tangent vector for $\boldsymbol{\psi}$, $\ell_{\beta\psi}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})[\boldsymbol{h}_1]$ be the derivative of $\ell_\beta(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ along the same path, $\ell_{\psi\psi}(\boldsymbol{\psi})[\boldsymbol{h}_1, \boldsymbol{h}_2]$ be the derivative of $\ell_\psi(\boldsymbol{\psi})[\boldsymbol{h}_1]$ along the path $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \epsilon \boldsymbol{h}_2$, with $\boldsymbol{h}_2$ being a tangent vector for $\boldsymbol{\psi}$, and $\ell_\xi(\boldsymbol{\xi})[\boldsymbol{h}_3]$ be the derivative of $R \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \{ S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}} \tilde{\boldsymbol{Z}} \}^2 / 2$ along the path $\boldsymbol{\xi} = \boldsymbol{\xi}^* + \epsilon \boldsymbol{h}_3$, with $\boldsymbol{h}_3$ being a tangent vector for $\boldsymbol{\xi}$. Let $\mathbb{P}_n$ and P denote the empirical and true probability measures, respectively. We impose the following conditions.

(C1) For $l = 1, \ldots, L$, g_l^* and $\boldsymbol{\eta}_l^*$ are unique, and g_l^* has bounded fourth derivative.

(C2) The support of $\boldsymbol{Z}$ is bounded, and $\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}$ has a bounded continuous support. Conditional on $\boldsymbol{Z}$, S has finite second moment.

(C3) The number of knots of the B-spline functions is such that $K_n^6 n^{-1/2} \to 0$ and $K_n^7 n^{-1/2} \to \infty$ as $n \to \infty$.

(C4) At $\beta = 0$, $\| \hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0 \| = o_p(n^{-1/4})$ for a suitable norm, and the estimator $\hat{\boldsymbol{\gamma}}$ satisfies

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = \mathbb{P}_n \boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0) + o_p(n^{-1/2}),$$

where $\boldsymbol{\ell}_\gamma^*$ is the efficient score function of $\boldsymbol{\gamma}$, such that $P\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0) = \mathbf{0}$, and $P\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)^{\mathrm{T}}$ is non-zero and finite.

(C5) The functions $\ell_\beta^2(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma}_0)$, $\ell_\psi^2(\boldsymbol{\psi})[\boldsymbol{h}_1]$, $\ell_{\beta\psi}(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma}_0)[\boldsymbol{h}_1]$, and $\ell_{\psi\psi}(\boldsymbol{\psi})[\boldsymbol{h}_1, \boldsymbol{h}_2]$ are Donsker for $(\boldsymbol{\psi}, \boldsymbol{\xi})$ belonging to a neighborhood of $(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)$ and $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ belonging to a bounded subset

of a suitable metric space. In addition, the information operator for the phenotype model $P\ell_{\psi\psi}(\boldsymbol{\psi_0})[\cdot,\cdot]$ is invertible under the null hypothesis H_0.

Remark 3.4. Conditions (C1) and (C2) pertain to regularity conditions on the variable with missing values and covariates. For g_l^* and $\boldsymbol{\eta}_l^*$ to be unique, we require that $\tilde{\boldsymbol{Z}}$ does not include $\tilde{\boldsymbol{X}}$ and that $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$ cannot be expressed as a function of linear terms of $\tilde{\boldsymbol{Z}}$. In practice, we let $\tilde{\boldsymbol{Z}}$ be a linear combination of the components of $\boldsymbol{Z}$ not present in $\tilde{\boldsymbol{X}}$, such that $\sum_{i=1}^n \tilde{\boldsymbol{Z}}_i \boldsymbol{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}} = \boldsymbol{0}$. Condition (C3) pertains to the rate at which the number of knots of the B-spline functions increases to infinity; particularly, the condition is satisfied with $K_n = O(n^{1/13})$. Conditions (C4) and (C5) are regularity conditions on the phenotype model, which are satisfied for common models, such as generalized linear models and proportional hazards models. For parametric models and the Cox proportional hazards model, the norm in condition (C4) is the Euclidean norm and the $\ell^\infty[0,t^*]$-norm, respectively, where t^* is the end of the study, and the metric space for $(\boldsymbol{h}_1,\boldsymbol{h}_2)$ in condition (C5) is the Euclidean space and the space of functions of bounded variation, respectively.

The asymptotic distribution of the robust imputation score statistic is given in the following theorem.

Theorem 3.1. *Under conditions (C1)-(C5) and $\beta = 0$, $U_\beta^{\mathrm{rob}}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\gamma}})$ is asymptotically zero-mean normal with variance*

$$V = P[\{\ell_\beta(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0) - \ell_\psi(\boldsymbol{\psi}_0)[\boldsymbol{h}_\psi] - \ell_\xi(\boldsymbol{\xi}^*)[\boldsymbol{h}_\xi] - \boldsymbol{I}_\gamma(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*)^{\mathrm{T}}\boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0)\}^2],$$

where $\boldsymbol{h}_\psi$ solves $P\ell_{\beta\psi}(\boldsymbol{\psi}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)[\cdot] = P\ell_{\psi\psi}(\boldsymbol{\psi}_0)[\boldsymbol{h}_\psi,\cdot]$, $\boldsymbol{h}_\xi = (h_{g,1}\ldots,h_{g,L},\boldsymbol{h}_{\eta,1},\ldots,\boldsymbol{h}_{\eta,L})$, such that $\boldsymbol{h}_{\eta,l} = \boldsymbol{0}$ and

$$h_{g,l}(t) = \frac{\mathrm{E}\{(1-R)I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)A(Y,\boldsymbol{X};\boldsymbol{\psi}_0) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l\}}{\mathrm{E}\{RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l\}}$$

for $l = 1,\ldots,L$,

$$\boldsymbol{I}_\gamma(\boldsymbol{\gamma},\boldsymbol{\xi}) = \sum_{l=1}^{L} \mathrm{E}[I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\boldsymbol{X}\{Rg_l'(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})h_{g,l}(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}) - (1-R)A(Y,\boldsymbol{X};\boldsymbol{\psi})g_l{}'(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})\}],$$

and f' denotes the first derivative of f for any function f.

Remark 3.5. The second and third terms in V are projections of the score function of $(\boldsymbol{\psi},\boldsymbol{\xi})$, and

$\boldsymbol{h}_\psi$ is the least-favorable direction of $\boldsymbol{\psi}$ for the phenotype model if the imputation model is assumed to be known. The fourth term in V is present because $\hat{\boldsymbol{\gamma}}$, instead of the true value, is used in the imputation model. The estimator $\hat{\boldsymbol{\gamma}}$ affects the imputation both by directly entering the imputation function $\hat{g}_l(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X})$ and by involving in the estimation of $\hat{g}_l$.

Motivated by Theorem 3.1, we propose an empirical variance estimator of the score statistic

$$\hat{V} = n^{-1}\sum_{i=1}^{n}[\{\ell_{\beta,i}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\gamma}}) - \ell_{\psi,i}(\hat{\boldsymbol{\psi}})[\hat{\boldsymbol{h}}_\psi] - \ell_{\xi,i}(\hat{\boldsymbol{\xi}})[\hat{\boldsymbol{h}}_\xi] - \hat{\boldsymbol{I}}_\gamma(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})\hat{\boldsymbol{\ell}}^*_\gamma(\hat{\boldsymbol{\psi}})\} - M]^2,$$

where $(\ell_{\beta,i},\ell_{\psi,i},\ell_{\xi,i})$ is $(\ell_\beta,\ell_\psi,\ell_\xi)$ evaluated at the observations of the ith subject, M is the sample mean of the first term in $\hat{V}$, and $(\hat{\boldsymbol{h}}_\psi,\hat{\boldsymbol{h}}_\xi,\hat{\boldsymbol{I}}_\gamma,\hat{\boldsymbol{\ell}}^*_\gamma)$ is the empirical version of $(\boldsymbol{h}_\psi,\boldsymbol{h}_\xi,\boldsymbol{I}_\gamma,\boldsymbol{\ell}^*_\gamma)$ evaluated at $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})$. Specifically, $\hat{\boldsymbol{h}}_\xi$ is obtained by performing the usual linear expansion of $\hat{\boldsymbol{\xi}}$ at $\boldsymbol{\xi}^*$, with the imputation model treated as a linear model with covariates $I(\tilde{\boldsymbol{X}}=\tilde{\boldsymbol{x}}_l)(\boldsymbol{B}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}},\tilde{\boldsymbol{Z}}^{\mathrm{T}})^{\mathrm{T}}$. The explicit form of $\hat{\boldsymbol{h}}_\xi$ is given in the proof of Theorem 3.2 in Chapter 3.6.1. We formulate the variance estimator under the linear model, the logistic model, and the Cox proportional hazards model in Chapter 3.6.2. The resulting score test statistic is $U^{\mathrm{rob}}_\beta(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\gamma}})^2/\hat{V}$. The validity of the robust score test is stated below.

Theorem 3.2. *Under conditions (C1)-(C5) and $\beta = 0$, the empirical variance estimator $\hat{V}$ converges almost surely to the true variance V, and the test statistic $U^{\mathrm{rob}}_\beta(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\gamma}})^2/\hat{V}$ converges in distribution to the chi-square distribution with one degree of freedom.*

Remark 3.6. The empirical variance estimator is consistent regardless of the missing-data mechanism and the imputation model. By contrast, the standard model-based variance estimator with imputed data is generally biased if the missing-data mechanism depends on the phenotype. The bias of the standard variance estimator under generalized linear models is derived in Chapter 3.6.3.

When the missing-data mechanism does not depend on the phenotype, the score statistic is unbiased under any imputation schemes; this result follows from the proof of Proposition 3.1. In this case, the proposed test is not required for bias correction, and one may wonder whether the inclusion of the B-spline terms and the stratification may lead to power loss; however, for the linear model, the asymptotic power of the proposed test is higher than or equal to that of the score test without the B-spline terms or stratification. The comparison of power between the proposed test

and the standard score test is difficult under more general settings because the power generally depends on the missing-data mechanism and high moments of S. The derivation for the power of the proposed test is given in Chapter 3.6.4.

3.3 Simulation Studies

Let $\boldsymbol{X} = (X_1, X_2, X_3)^{\mathrm{T}}$, where X_1, X_2, and X_3 are independent standard normal, Bernoulli(0.5), and Binomial$(2, 0.25)$, respectively. Let $\boldsymbol{G}$ be a vector of other covariates that are used to predict S. In particular, $\boldsymbol{G} = (G_1, \ldots, G_4)$, where G_j $(j = 1, \ldots, 4)$ is independent Binomial$(2, 0.3)$. In cancer genomics, X_1, X_2, and X_3 may represent (standardized) age, gender, and tumor stage, respectively, and $\boldsymbol{G}$ may represent genotypes at four loci. We generated the phenotype Y using the linear predictor $r(S, \boldsymbol{X}) = \gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X} + \beta S$ under the linear, logistic, and proportional hazards models. For all models, we set $\boldsymbol{\gamma} = (1, -1, 0.5)^{\mathrm{T}}$. For the linear model, we generated $Y \sim \mathrm{N}\{r(S, \boldsymbol{X}), 1\}$ with $\gamma_0 = 0$. For the logistic model, we set $\mathrm{logit}^{-1}\{P(Y = 1 \mid S, \boldsymbol{X})\} = r(S, \boldsymbol{X})$, where γ_0 was chosen such that $P(Y = 1) \approx 0.15$. For the proportional hazards model, we generated Y with the hazard function $\lambda(t \mid S, \boldsymbol{X}) = 0.5te^{r(S,\boldsymbol{X})}$ and $\gamma_0 = 0$. The censoring variable was generated independently from Unif$(0, \tau)$, where τ was chosen such that the censoring proportion was about 40%. We considered two models for S: with Model 1, $S = X_1 + X_2 + 0.3X_3 + 0.4(G_1 - G_2 + G_3 - G_4) + \mathrm{N}(0, 1)$; and with Model 2, $S = (X_1 + X_2) + 0.1(X_1 + X_2)^2 + 0.3I(X_3 = 2) + 0.4(G_1 - G_2 + G_3 - G_4) + \mathrm{N}(0, 1)$.

We compared the performance of six tests: (1) the standard score test using complete data only; (2) the standard score test with missing values imputed under a linear model of S on $\boldsymbol{Z} \equiv (\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}$; (3) Lawless (2016)'s score test based on the same model of S as (2); (4) Hu et al. (2015)'s score test with the imputed data of (2); (5) the proposed score test with stratification variable $\tilde{X} = X_2$ and $\tilde{\boldsymbol{Z}}$ being that specified in Remark 3; and (6) the imputation score test with missing data imputed using a linear model of S on $\boldsymbol{Z} = (\boldsymbol{X}^{\mathrm{T}}, X_1^2, X_1X_2, I(X_3 = 2), \boldsymbol{G}^{\mathrm{T}})^{\mathrm{T}}$ and the empirical variance estimator. We refer to methods (1)-(6) as the complete-case analysis, the simple imputation method, Lawless' method, Hu's method, the proposed imputation method, and the full-model imputation method, respectively. The last method is the gold standard but is not practical because it requires correct specification of a complex missing-data model. Derkach et al. (2015)'s method was not included because it requires the covariates in the imputation model to be discrete and is identical to Lawless (2016)'s method when a linear imputation model is assumed. Note that the missing-data models used by all the methods are correct under Model 1, but only the missing-data model used by the

full-model imputation method is correct under Model 2. For the proposed imputation method, we chose the degree and number of knots of the B-spline functions using five-fold cross-validation separately for each stratum. For the lth stratum, the grid point τ_k ($k = 1, \ldots, K_n - m - 2$) was set to be the empirical $k/(K_n - m + 1)$-quantile of $\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i$ among subjects with $R_i = 1$ and $\boldsymbol{X}_i = \tilde{\boldsymbol{x}}_l$, $\tau_0 = \min_{\tilde{\boldsymbol{X}}_i = \tilde{x}_l} \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i$, and $\tau_{K_n - m - 1} = \max_{\tilde{\boldsymbol{X}}_i = \tilde{x}_l} \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}_i$. Lawless' and Hu's methods are not applicable to the survival phenotype.

We considered two missing-data mechanisms. Mechanism 1 is missing completely at random, where the missing-data status is independent of other variables. For Mechanism 2, the missing-data status was generated separately for two subsets of subjects: one subset consisted of all subjects with $X_2 = 1$, and a random sample of subjects from the subset were selected for observation of S; the other subset consisted of all subjects with $X_2 = 0$, and subjects from the subset were selected for observation of S based on the phenotype. For the continuous and survival phenotypes, an equal number of subjects at the two extreme tails of the phenotype distribution were selected. For the binary phenotype, all subjects with $Y = 1$ were selected, and sufficient subjects with $Y = 0$ were selected such that the desired missing proportion was attained. The missing proportion was set to be the same between the two subsets of subjects. This setting mimics a study where two datasets with different sampling schemes are combined.

We considered a sample size of 1,500 and missing proportions ranging from 30% to 60%. For each setting, we simulated 1,000,000 and 100,000 replicates for $\beta = 0$ and $\beta \neq 0$, respectively. The nominal significance level was set to 10^{-3}. We plot the rejection probability against the missing proportion for the two models of S and the two missing-data mechanisms; see Figures 3.1–3 for the results of the linear, logistic, and Cox proportional hazards models, respectively.

Under Mechanism 1, all methods have correct type I error. Under Model 1 and Mechanism 2, the simple imputation method has inflated type I error because the variance of the score statistic is underestimated. The complete-case analysis is also invalid except for the binary phenotype, but the type I error inflation is not as severe; the complete-case analysis for the binary phenotype has correct type I error because of the special structure of the logistic model (Prentice and Pyke 1979). Hu et al. (2015)'s variance estimator requires that both the actual and imputed S variables are independent of $\boldsymbol{X}$, which does not hold under either Model 1 or Model 2. As a result, the variance is overestimated under Mechanism 2, which leads to type I error deflation. The remaining methods

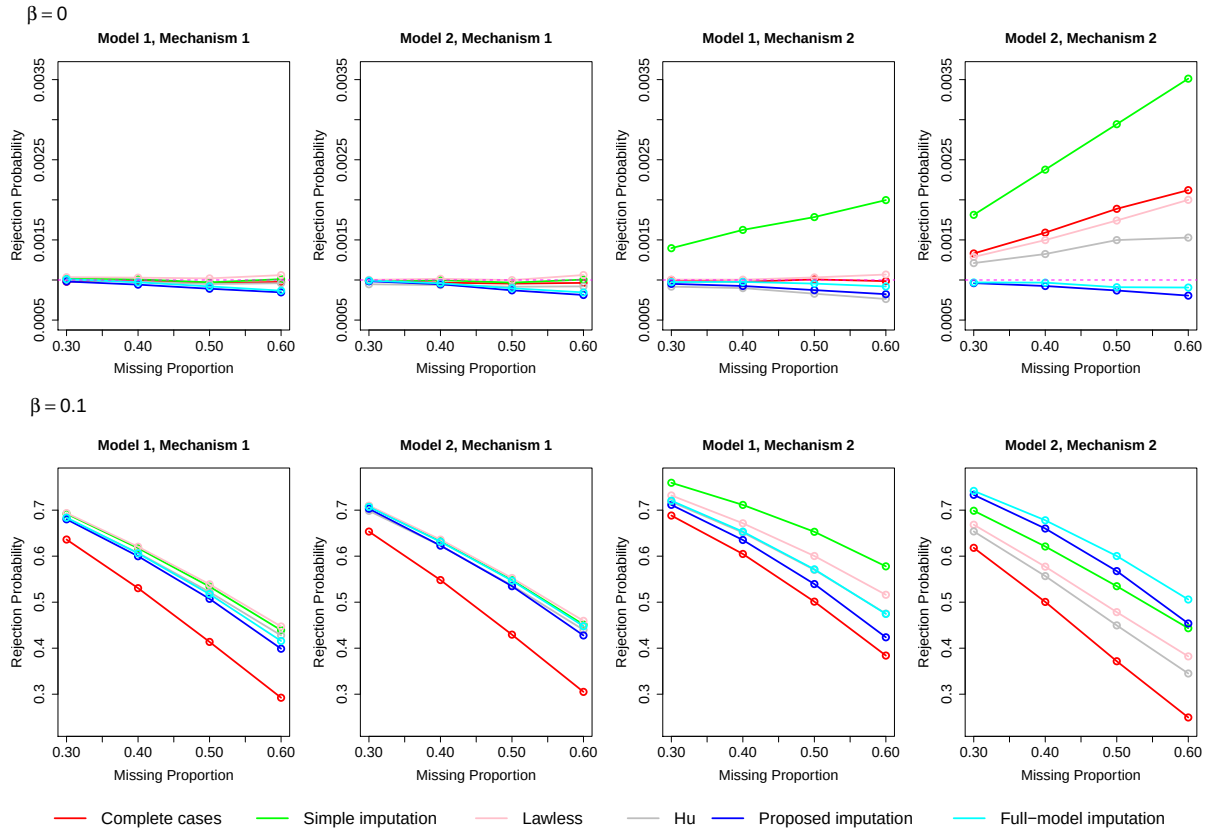


Figure 3.1. Rejection Probabilities Under the Null and Alternative Hypotheses for the Continuous Phenotype.

have consistent variance estimators and, therefore, have correct type I error. Under Model 2 and Mechanism 2, the score statistics of the complete-case analysis and the methods based on a model of S on linear terms of $\boldsymbol{X}$ are generally biased, giving rise to type I error inflation in most cases. Hu's method exhibits type I error deflation under the logistic model because the bias of the score statistic is offset by the overestimation of the variance in this specific setting. (Because the absolute bias of the score statistic tends to infinity as $n \to \infty$, Hu's method would yield type I error inflation for large enough sample size.) The proposed imputation method is valid even though the imputation model is misspecified because the score statistic is unbiased. The full-model imputation method is also valid because the imputation model is correct.

The power of the complete-case analysis is generally low because it discards useful information. Under Model 1 or Mechanism 1, all valid methods that use the whole dataset have similar power. Under Model 2 and Mechanism 2, the full-model imputation method is the most powerful among

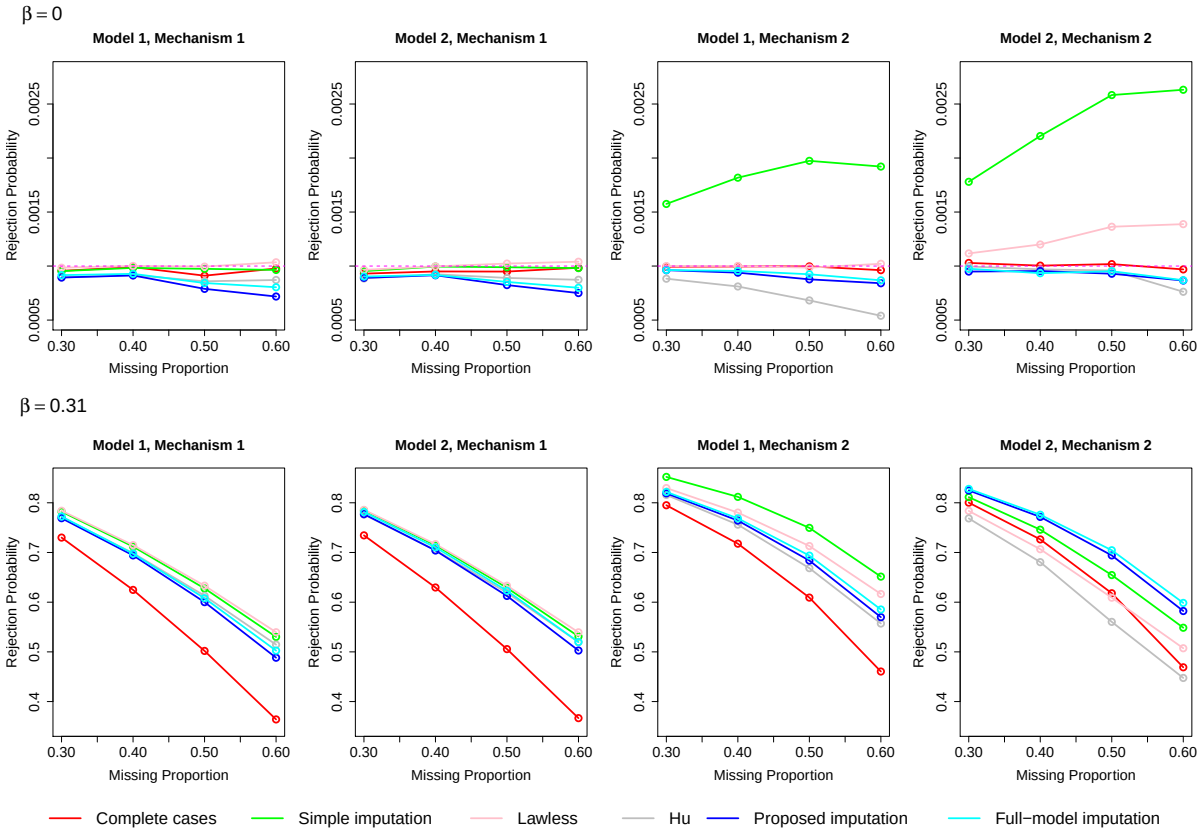


Figure 3.2. Rejection Probabilities Under the Null and Alternative Hypotheses for the Binary Phenotype.

the valid methods because a correct imputation model is assumed. However, this method cannot be used in practice because it requires knowledge of the true relationship between S and $\boldsymbol{Z}$. The proposed imputation method is only slightly less powerful than the full-model imputation method. The bias of the score statistic of the other methods can lead to substantially low power.

3.4 Real Data Analysis

We analyzed a dataset of patients with serous ovarian cancer from TCGA (The Cancer Genome Atlas Research Network 2011). In the study, most subjects had available genomic data, including data on DNA copy number, somatic mutation, and levels of expression of mRNA measured by microarray platforms. Only a subset of subjects had enough tissue sample left for RNA sequencing, which was introduced after the study had begun. Demographic and clinical variables, including age at diagnosis, tumor stage, tumor grade, time to tumor progression, and time to death, were available for most subjects. The median follow-up time was about 2.5 years, and roughly 30% of

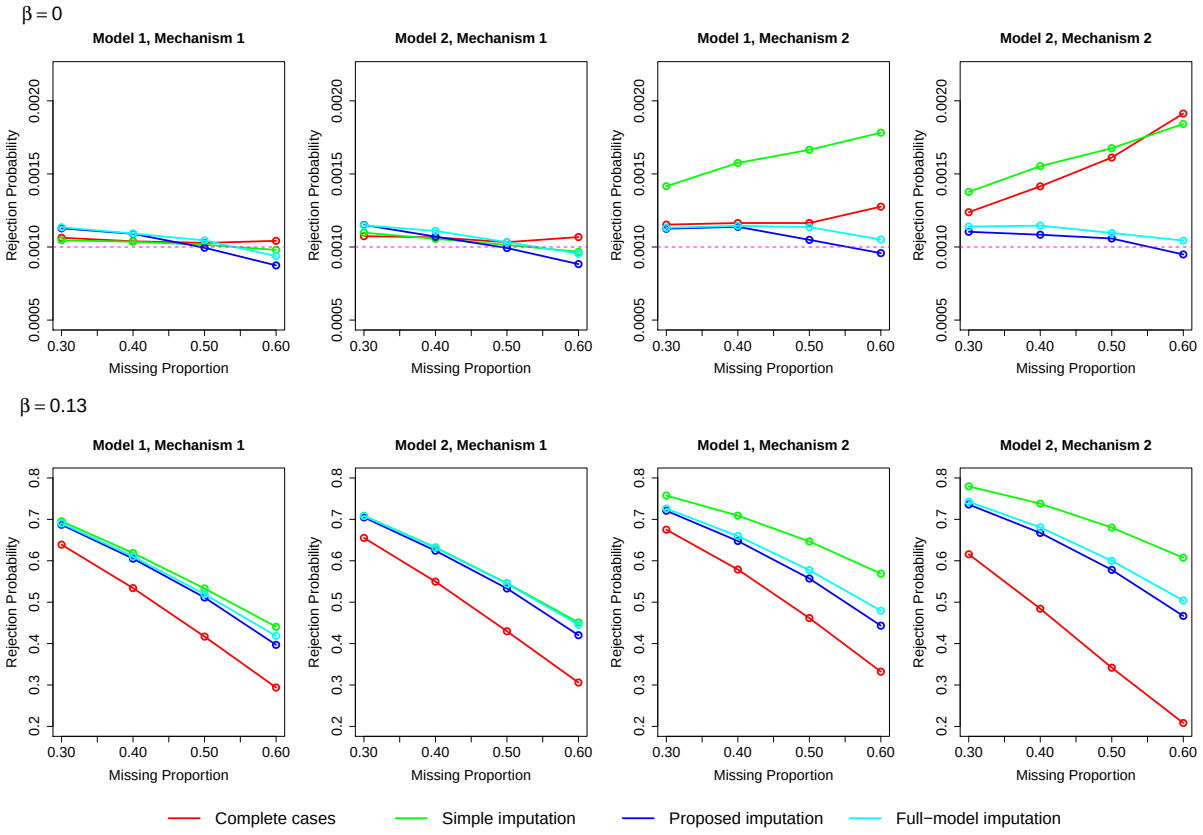


Figure 3.3. Rejection Probabilities Under the Null and Alternative Hypotheses for the Survival Phenotype.

the patients were lost to follow-up before tumor progression or death. The data are available at http://gdac.broadinstitute.org/.

We focused on testing the association between mRNA expression, measured by RNA sequencing, and progression-free survival time. We used the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values for the mRNA expression variable. The number of transcripts with RNA sequencing data was about 57,000. We considered a subset of 9,068 genes that were mutated in samples from more than five subjects. The number of subjects with available mutation, copy number, and clinical data was 407, approximately 30% of whom did not have RNA sequencing data.

We fit the Cox model for progression-free survival and included age, tumor stage, tumor grade, the squared term of age, the interaction between age and (dichotomized) tumor stage, and the interaction between age and (dichotomized) tumor grade as covariates. The predictors in the imputation model of the RNA sequencing data included the covariates in the phenotype model,

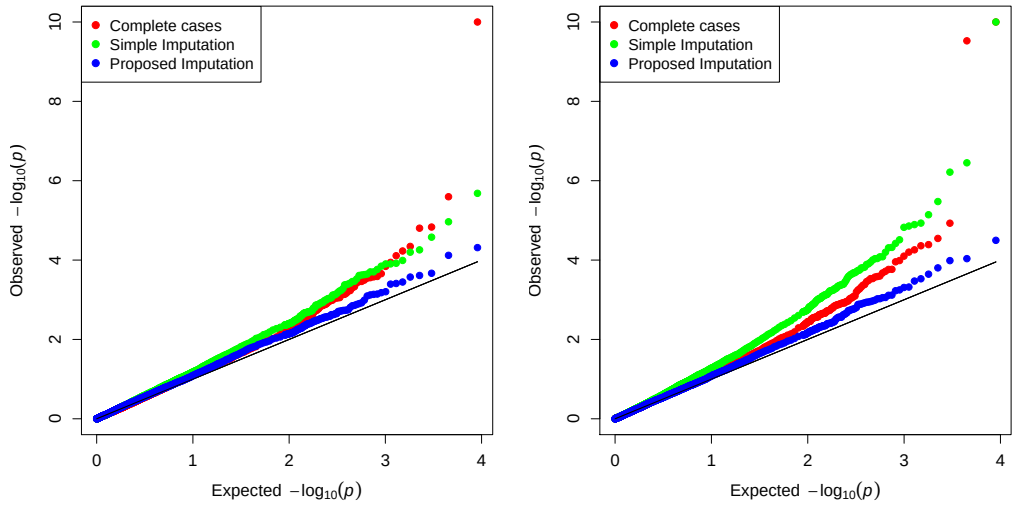


Figure 3.4. Quantile-Quantile Plots for the RNA-Seq Analysis of the TCGA Ovarian Cancer Data. The left plot shows the results for the original data, and the right plot shows the results with the missing proportion increased to 60%. The p-values are truncated at 10^{-10}.

together with somatic mutation, copy number, and mRNA microarray expression. Microarray mRNA expression and somatic mutation were excluded from the imputation model if they were missing or too sparse. We did not include a stratification variable. The B-spline functions were selected in the same way as in the simulation studies. For comparison, we performed the standard score test with only the complete cases and with the missing values imputed under a linear model. For further illustration, we performed the proposed test and the standard score test on the dataset with the missing proportion increased to 60%, where the RNA sequencing variables for subjects with intermediate survival or censoring time were treated as missing. The quantile-quantile plots are shown in Figure 3.4.

For the original dataset, the p-values from the proposed method agree with the expectation that most gene expressions are not associated with progression-free survival. The complete-case analysis and the simple imputation method yielded excessive false-positive signals because the standard variance estimates of the score statistic are smaller than the empirical variance. With extra missing data, the inflation of type I error is more severe for the simple imputation method, whereas the type I error is preserved by the proposed method.

The top ten genes identified by the proposed method from the original data are presented

Table 3.1. Top Genes and Their p-values in the RNA-Seq Analysis of the TCGA Ovarian Cancer Data.

Gene	Proposed method	Complete cases	Single imputation
WDR91	4.85E-05	3.20E-04	2.65E-05
SLC4A8	7.60E-05	7.77E-05	1.08E-05
NPAS3	2.15E-04	9.11E-04	1.98E-04
PLAUR	2.44E-04	5.38E-03	9.40E-04
TGFBI	2.68E-04	5.90E-05	6.32E-05
ST3GAL1	3.57E-04	9.57E-03	3.30E-03
LRIG2	3.88E-04	4.31E-03	1.42E-04
DUSP1	4.03E-04	1.26E-03	2.63E-03
GALNT6	6.27E-04	1.87E-03	2.25E-04
VMO1	6.63E-04	9.06E-02	4.24E-03

in Table 3.1. Several of them have been reported to be related to ovarian cancer (Denkert et al. 2002; Wang et al. 2005; Ahmed et al. 2007; Ween et al. 2012; Arend et al. 2013; Wang et al. 2014; Killeen et al. 2014; Caburet et al. 2015). Among those genes, the associations between progression-free survival time and the expressions of WDR91, SLC4A8, NPAS3, PLAUR, ST3GAL1, LRIG2, DUSP1, GALNT6, and VMO1 are more significant under the proposed method than under the complete-case analysis. The significance levels for associations between progression-free survival time and expressions of of PLAUR, ST3GAL1, DUSP1, and VMO1 are lower under the simple imputation method than the proposed method.

3.5 Discussion

In this chapter, we propose a robust score test for the association between a phenotype and a genomic variable with partially missing values. The test is based on a semiparametric model for the genomic variable, where the semiparametric component ensures that under the null hypothesis, the score statistic with imputed values is unbiased for general missing-data mechanisms and arbitrary distributions of the genomic variable. Because the nonparametric function in the imputation model is one-dimensional regardless of the dimension of the covariates, the score test is computationally feasible with a large number of covariates. In addition to correcting for the bias of the score statistic, the semiparametric component results in a better fit of the imputation model, which leads to power gain even when data are missing completely at random. When the missing-data mechanism depends on the phenotype, the proposed test has correct type I error, whereas the standard score test is generally invalid. When the missing-data mechanism is independent of the phenotype, both the

proposed and standard score tests have correct type I error, but the proposed test is asymptotically more powerful.

The validity of the proposed test follows from two special properties of the score statistic under the null hypothesis. First, the phenotype model does not involve the variable with missing values, and the score statistic derived under the full likelihood coincides with the imputation score statistic. Second, the score statistic is mean zero if the expectations of the actual and imputed values conditional on a low-dimensional function of the covariates are equal. As a result, single imputation yields a valid score statistic if the expectation of the variable with missing values conditional on the low-dimensional function of the covariates is correctly specified. These two properties do not hold under the alternative hypothesis, making parameter estimation with missing data a much more challenging problem than hypothesis testing. For estimation, single imputation generally yields underestimation of standard errors (Little 1992), and a correct specification of the missing-data model is required for valid inference.

Our work can be extended in several directions. First, we have focused on a continuous genomic variable that is either exactly observed or missing. We may allow for a binary or categorical variable by incorporating the proposed semiparametric component into a generalized linear modeling framework. We may also consider genomic variables that are subject to censoring or detection limits, as in the case of metabolomics data (Yu et al. 2014). In this case, the conditional mean of the genomic variable cannot be consistently estimated using simple least-squares estimation, and additional assumptions on the distribution of the genomic variable are necessary.

Second, it would be of interest to perform a joint test for multiple genomic variables, where each variable may have a separate pattern of missing values. This extension can be applied to many existing testing procedures that involve a multivariate score statistic, such as the sequence kernel association test for rare variants (Wu et al. 2011), tests for meta-analysis of sequencing data (Tang and Lin 2013), and the joint test for multiple genomic variables (Huang et al. 2014). Joint modeling of multiple variables with missing values is more challenging than modeling a single variable with missing value, when the pattern of the missing values for each variable do not overlap. Nevertheless, fitting a separate imputation model for each variable is not preferable, as it results in efficiency loss when the variables are correlated.

Finally, we have assumed that the missing-data mechanism depends only on the phenotype and

a set of discrete covariates. The methodology can be extended to allow the missing-data mechanism to depend on continuous covariates by relating the variable with missing values to a nonparametric function of the phenotype and covariates. In addition, we may consider a missing-data mechanism that depends on a different phenotype; this scenario is common in the analysis of secondary phenotypes. In this case, the function through which the covariates affect the alternative phenotype must be estimated, and the variable with missing values should be modeled nonparametrically on that function.

3.6 Technical Details and Additional Results

3.6.1 Proofs of Technical Results

In this section, we prove the two propositions and the two theorems put forth in Chapter 3.2.

Proof of Proposition 3.1. The expected value of $U_\beta^{\text{imp}}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)$ is

$$\begin{aligned}
& \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)\{RS + (1-R)\tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\}] \\
= \; & \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R\{S - \tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\}] + \mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)\tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*)\} \\
= \; & \mathrm{E}[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R\mathrm{E}\{S - \tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid Y, \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}\}] \\
= \; & \mathrm{E}[\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}\}\mathrm{E}\{S - \tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}\}],
\end{aligned}$$

where the second and third equalities follow from the facts that Y is independent of $\boldsymbol{Z}$ conditional on $\boldsymbol{X}$ and that R is independent of $(S, \boldsymbol{Z})$ conditional on $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}})$, respectively. Clearly, the above expectation is zero since $\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \mathrm{E}\{\tilde{S}(\boldsymbol{Z}; \boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}\}$. □

Proof of Proposition 3.2. Because S is independent of R conditional on $(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}})$ under $\beta = 0$,

$$\begin{aligned}
& \mathrm{E}[R\{S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}^2 | \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l] \\
= \; & \mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\mathrm{E}[\{S - g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}^2 | \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l] \\
= \; & \mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\mathrm{E}([S - \{g_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}}\mathrm{E}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\} \\
& \quad -\boldsymbol{\eta}_l^{\mathrm{T}}\{\tilde{\boldsymbol{Z}} - \mathrm{E}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\}]^2 | \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l).
\end{aligned}$$

Therefore, $(g_l^*, \boldsymbol{\eta}_l^*)$ satisfies $g_l^*(t) + \boldsymbol{\eta}_l^{*\mathrm{T}}\mathrm{E}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) = \mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X} = t, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)$. The

mean of the imputed S is

$$\begin{aligned}
\mathrm{E}\{\tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}\} &= \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + \boldsymbol{\eta}_l^{*\mathrm{T}}\mathrm{E}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\} \\
&= \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \\
&= \mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}).
\end{aligned}$$

The desired result follows from Proposition 3.1. □

Before proving Theorem 3.1, we present the following lemma, which pertains to the consistency of $\hat{\boldsymbol{\xi}}$.

Lemma 3.1. *Under conditions (C1)-(C4), $\|\hat{g}_l - g_l^*\|_{W^{1,\infty}} \to 0$, and $\|\hat{g}_l - g_l^*\|_{L_2(X)}^2 + \|\hat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*\|_2^2 = o_p(n^{-1/2})$ for $l = 1, \ldots, L$, where for any function h that has bounded first derivative, $\|h\|_{W^{1,\infty}} = \|h\|_{\ell^\infty} + \|h'\|_{\ell^\infty}$, and $\|h\|_{L_2(X)} = [\mathrm{E}\{h(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})^2\}]^{1/2}$.*

Proof of Lemma 3.1. By Theorem 6.25 of Schumaker (2007), there exists $\tilde{g}_l = \sum_{k=1}^{K_n} \tilde{\alpha}_{lk} B_k$, such that $\|\tilde{g}_l - g_l^*\|_{W^{1,\infty}} \le O(K_n^{-2})$ and $\|\tilde{g}_l - g_l^*\|_{L_2} \le O(K_n^{-7/2})$ for each $l = 1, \ldots, L$. Let $\phi_l(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma}) = -RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{S - g_l(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}^2/2$. By the definition of $(\hat{g}_l, \hat{\boldsymbol{\eta}}_l)$, $\mathbb{P}_n\phi_l(\hat{g}_l, \hat{\boldsymbol{\eta}}_l; \hat{\boldsymbol{\gamma}}) \ge \mathbb{P}_n\phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}})$. By the concavity of ϕ_l,

$$\mathbb{P}_n\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}}) \ge \mathbb{P}_n\phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}}),$$

where $\epsilon = (1 + \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}| + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|)^{-1}$. Therefore,

$$\begin{aligned}
& (\mathbb{P}_n - P)\{\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}}) - \psi(\tilde{g}_l, \boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}})\} \\
\ge\ & -P\{\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}}) - \phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}})\}. \qquad (3.1)
\end{aligned}$$

Note that

$$\begin{aligned}
\phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}}) - \phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) &= R\Big\{S - \sum_{k=1}^{K_n} \tilde{\alpha}_{lk} B_k(\tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{*\mathrm{T}}\tilde{\boldsymbol{Z}}\Big\} \sum_{k=1}^{K_n} \tilde{\alpha}_{lk} B_k'(\tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{X})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^{\mathrm{T}}\boldsymbol{X} \\
&\le O_p(|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|) \max_k \|B_k'\|_{\ell^\infty} \\
&\le O_p(K_n n^{-1/2}),
\end{aligned}$$

where the first inequality follows because $\tilde{\alpha}_{lk}$ is bounded, and $\tilde{\boldsymbol{\gamma}}$ is some value between $\boldsymbol{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}$. Likewise, because

$$\epsilon\hat{\alpha}_{lk} + (1-\epsilon)\tilde{\alpha}_{lk} = \frac{\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}}{c_0 + |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|} + \frac{1 + |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|}{c_0 + |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|}\tilde{\alpha}_{lk}$$

for some $c_0 > 1$,

$$\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \hat{\boldsymbol{\gamma}}) - \phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) \leq O_p(K_n n^{-1/2}).$$

Therefore, (3.1) implies that

$$\begin{aligned} & (\mathbb{P}_n - P)\{\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\} + O_p(K_n n^{-1/2}) \\ \geq \; & -P\{\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\}. \end{aligned} \tag{3.2}$$

By Theorem 2.6.15 of van der Vaart and Wellner (1996), referred to as VW hereafter, $\{\{\boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}} + \sum_{k=1}^{K_n} \alpha_{lk} B_k(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\}^2 : \max_{k,j}(|\alpha_{lk}|, |\eta_{lj}|) < M\}$ is a Vapnik-Chervonenkis (VC) class with VC index at most $(K_n + q)^2 + 2$ for any $M < \infty$. By Theorem 2.6.7 of VW, we show that

$$N(c_1\|F\|_2, \mathcal{F}, L_2) \leq c_2\{(K_n+q)^2 + 2\}(16e)^{\{(K_n+q)^2+2\}^2}\left(\frac{1}{c_1}\right)^{2\{(K_n+q)^2+1\}}$$

for any $0 \leq c_1 \leq 1$, where N is the covering number, c_2 is a constant, $\mathcal{F} = \{\phi_l(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma}_0) : g_l(\cdot) = \sum_{k=1}^{K_n} \alpha_{lk} B_k(\cdot), \max_{k,j}(|\alpha_{lk}|, |\eta_{lj}|) < M\}$, and F is an envelope function of $\mathcal{F}$ that consists of second-order terms of $(S, \tilde{\boldsymbol{Z}})$. Therefore, the uniform entropy of the class of functions is $O(K_n^2)$. By Theorem 2.14.1 of VW and Markov's inequality, the left-hand side of (3.2) is $O_p(K_n n^{-1/2})$.

By the linear expansion of ϕ_l at g_l^*, $P\phi_l(\tilde{g}_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)$ is equal to

$$\begin{aligned} & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \frac{1}{2}\sum_{l=1}^{L} P(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\mathrm{E}[\mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{\tilde{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\}^2 \mid \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l] \\ \leq \; & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) + O_p(\|\tilde{g}_l - g_l^*\|_{L_2(X)}^2) \\ \leq \; & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) + O_p(K_n^{-7}), \end{aligned}$$

where the first equality above follows because R and S are conditionally independent given $\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}$

and $\tilde{\boldsymbol{X}}$. Likewise,

$$
\begin{aligned}
& P\phi_l(\epsilon\hat{g}_l + (1-\epsilon)\tilde{g}_l, \epsilon\hat{\boldsymbol{\eta}}_l + (1-\epsilon)\boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) \\
= \; & P\phi_l(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) - \frac{1}{2}\sum_{l=1}^{L} P(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \\
& \times \mathrm{E}\{\mathrm{E}(R \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)[\epsilon\{\hat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - \tilde{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + (\hat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*)^{\mathrm{T}}\tilde{\boldsymbol{Z}}\} + \{\tilde{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\}]^2\}.
\end{aligned}
$$

We conclude that

$$
\epsilon^2\Big(\|\hat{g}_l - \tilde{g}_l\|_{L_2(X)}^2 + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|^2 \Big) \leq O_p(K_n n^{-1/2}) + O_p(K_n^{-7}).
$$

Because $\|\hat{g}_l - \tilde{g}_l\|_{L_2(X)}^2 \geq c_3 \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|^2$ for some $c_3 > 0$,

$$
\begin{aligned}
& \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|^2 + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|^2 \\
\leq \; & \Big(1 + \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}| + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|\Big)^2 \{O_p(K_n n^{-1/2}) + O_p(K_n^{-7})\} \\
= \; & \Big(1 + \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|^2 + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|^2\Big) K_n^2 \{O_p(K_n n^{-1/2}) + O_p(K_n^{-7})\},
\end{aligned}
$$

which implies that $\sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|^2 + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|^2 \leq O_p(K_n^3 n^{-1/2}) + O_p(K_n^{-5})\}$. Note that

$$
\begin{aligned}
\|\hat{g}_l - \tilde{g}_l\|_{W^{1,\infty}} & \leq \max_x B_k(x) \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}| + \max_x |B_k{}'(x)| \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}| \\
& \leq O_p\Big(K_n \sum_{k=1}^{K_n} |\hat{\alpha}_{lk} - \tilde{\alpha}_{lk}|\Big) = O_p(K_n^3 n^{-1/4} + K_n^{-1}).
\end{aligned}
$$

Because $K_n^6 n^{-1/2} \to 0$, both $\|\hat{g}_l - g_l^*\|_{W^{1,\infty}}$ and $\sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|^2$ converge to zero in probability.

To establish the rate of convergence, we replace ϵ by 1 in (3.2). The left-hand side of the resulting inequality is $o_p(n^{-1/2})$ because $\phi_l(g_l, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)$ is Donsker in a neighborhood of g_l^*. By the linear expansion on the right side,

$$
o_p(n^{-1/2}) \geq O_p(K_n^{-7}) + \|\hat{g}_l - g_l^*\|_{L_2(X)}^2 + \sum_{j=1}^{q} |\hat{\eta}_{lj} - \eta_{lj}^*|^2.
$$

The desired rate of convergence follows since $K_n^7 n^{-1/2} \to \infty$. □

Proof of Theorem 3.1. The imputation score statistic is

$$\begin{aligned} & n^{1/2}\mathbb{P}_n \ell_\beta(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}}) \\ = \;& n^{1/2}\mathbb{P}_n \ell_\beta(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0) + n^{1/2} P \ell_\beta(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}}) + n^{1/2}(\mathbb{P}_n - P)\{\ell_\beta(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}}) - \ell_\beta(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)\} \\ = \;& n^{1/2}\mathbb{P}_n \ell_\beta(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0) + n^{1/2} P\{\ell_\beta(\hat{\boldsymbol{\psi}}, \boldsymbol{\xi}^*; \hat{\boldsymbol{\gamma}}) - \ell_\beta(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)\} \\ & + n^{1/2} P\{\ell_\beta(\boldsymbol{\psi}_0, \hat{\boldsymbol{\xi}}; \boldsymbol{\gamma}_0) - \ell_\beta(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)\} + o_p(1). \end{aligned} \tag{3.3}$$

The second equality above follows because the convergence rate of $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is at least $n^{-1/4}$ and $\ell_\beta(\boldsymbol{\psi}, \boldsymbol{\xi}; \boldsymbol{\gamma})$ is Donsker over a neighborhood of $(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*, \boldsymbol{\gamma}_0)$. By the linear expansion, the second term of (3.3) is equal to

$$n^{1/2}\mathrm{E}\Big\{A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)(1-R)\sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) g_0^{*\prime}(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}^{\mathrm{T}}\Big\}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + n^{1/2} P \ell_{\beta\psi}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)[\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0],$$

up to an $o_p(1)$ term. Let $\boldsymbol{I}_\gamma^{(1)}(\boldsymbol{\psi}, \boldsymbol{\xi}) = -\mathrm{E}\{A(Y, \boldsymbol{X}; \boldsymbol{\psi})(1-R)\sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) g_l'(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\}$. Clearly, the first term of the above expression is $-n^{1/2}\boldsymbol{I}_\gamma^{(1)}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)^{\mathrm{T}}\mathbb{P}_n \boldsymbol{\ell}_\gamma^*(\boldsymbol{\psi}_0) + o_p(1)$ by condition (C4). By condition (C5), $\boldsymbol{h}_\psi$ exists, and the second term of the above expression is equal to $n^{1/2}\mathbb{P}_n \ell_\psi(\boldsymbol{\psi}_0)[\boldsymbol{h}_\psi] + o_p(1)$.

The third term of (3.3) is equal to

$$n^{1/2}\mathrm{E}\Big[A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0)(1-R)\sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{\hat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}) + (\hat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_0^*)^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}\Big].$$

Let $\ell_{g,l}(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma})[h_{g,l}] \equiv RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{S - g_l(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}h_{g,l}(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X})$ and $\boldsymbol{\ell}_{\eta,l}(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma}) \equiv RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{S - g_l(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}) - \boldsymbol{\eta}_l^{\mathrm{T}}\tilde{\boldsymbol{Z}}\}\tilde{\boldsymbol{Z}}$ be the score function for g_l and $\boldsymbol{\eta}_l$, respectively. Let $\tilde{h}_{g,l}$ be the B-spline approximation of $h_{g,l}$ on the same grid as $\hat{g}_l$. By the definition of $\hat{g}_l$ and $\hat{\boldsymbol{\eta}}_l$,

$$(\mathbb{P}_n - P)\{\ell_{g,l}(\hat{g}_l, \hat{\boldsymbol{\eta}}_l; \hat{\boldsymbol{\gamma}})[\tilde{h}_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}}\boldsymbol{\ell}_{\eta,l}(\hat{g}_l, \hat{\boldsymbol{\eta}}_l; \hat{\boldsymbol{\gamma}})\} = -P\{\ell_{g,l}(\hat{g}_l, \hat{\boldsymbol{\eta}}_l; \hat{\boldsymbol{\gamma}})[\tilde{h}_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}}\boldsymbol{\ell}_{\eta,l}(\hat{g}_l, \hat{\boldsymbol{\eta}}_l; \hat{\boldsymbol{\gamma}})\}.$$

Because $\hat{g}_l$ and $\tilde{h}_{g,l}$ have bounded first derivatives and $\ell_{g,l}$ and $\boldsymbol{\ell}_{\eta,l}$ are differentiable, $\ell_{g,l}(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma})[h_{g,l}]$ and $\boldsymbol{\ell}_{\eta,l}(g_l, \boldsymbol{\eta}_l; \boldsymbol{\gamma})$ are Donsker in a neighborhood of $(g_l^*, \boldsymbol{\eta}_l^*, \boldsymbol{\gamma}_0, h_{g,l})$. By the Donsker property of

the $\ell_{g,l}$ and $\boldsymbol{\ell}_{\eta,l}$ and the consistency results of Lemma 3.1,

$$\begin{aligned}
&(\mathbb{P}_n - P)\{\ell_{g,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)[h_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \boldsymbol{\ell}_{\eta,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\} \\
= \;& \mathrm{E}[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) g_l^{*\prime}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X})\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\} \boldsymbol{X}^{\mathrm{T}}](\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \\
&+ \mathrm{E}[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}\{\hat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + (\hat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*)^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}] + o_p(n^{-1/2}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathrm{E}[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}\{\hat{g}_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) - g_l^*(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + (\hat{\boldsymbol{\eta}}_l - \boldsymbol{\eta}_l^*)^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}] \\
= \;& (\mathbb{P}_n - P)\{\ell_{g,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)[h_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \boldsymbol{\ell}_{\eta,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0) \\
&- \mathrm{E}[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) g_l^{*\prime}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X})\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\} \boldsymbol{X}^{\mathrm{T}}] \boldsymbol{\ell}_{\gamma}^*(\boldsymbol{\psi}_0)\} + o_p(n^{-1/2}).
\end{aligned}$$

It is easy to see that $h_{g,l}$ and $\boldsymbol{h}_{\eta,l}$ solve

$$\begin{aligned}
&\mathrm{E}[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{h_{g,l}(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\} w(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \tilde{\boldsymbol{Z}})] \\
= \;& \mathrm{E}[(1 - R) I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) A(Y, \boldsymbol{X}; \boldsymbol{\psi}_0) w(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}, \tilde{\boldsymbol{Z}})] \qquad (3.4)
\end{aligned}$$

for all $w \in \{w(t, \tilde{\boldsymbol{Z}}) = w_1(t) + \boldsymbol{w}_2^{\mathrm{T}} \tilde{\boldsymbol{Z}} : w_1 \text{ has bounded fourth derivatives}, \boldsymbol{w}_2 \in \mathbb{R}^q\}$. Therefore,

$$\begin{aligned}
&n^{1/2} P\{\ell_{\beta}(\boldsymbol{\psi}_0, \hat{\boldsymbol{\xi}}; \boldsymbol{\gamma}_0) - \ell_{\beta}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*; \boldsymbol{\gamma}_0)\} \\
= \;& n^{1/2} \mathbb{P}_n \Big[\sum_{l=1}^{L} \{\ell_{g,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)[h_{g,l}] + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \boldsymbol{\ell}_{\eta,l}(g_l^*, \boldsymbol{\eta}_l^*; \boldsymbol{\gamma}_0)\} - \boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\ell}_{\gamma}^*(\boldsymbol{\psi}_0) \Big] + o_p(1) \\
= \;& n^{1/2} \mathbb{P}_n \{\ell_{\xi}(\boldsymbol{\xi}^*)[\boldsymbol{h}_{\xi}] - \boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}_0, \boldsymbol{\xi}^*)^{\mathrm{T}} \boldsymbol{\ell}_{\gamma}^*(\boldsymbol{\psi}_0)\} + o_p(1),
\end{aligned}$$

where $\boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \sum_{l=1}^{L} \mathrm{E}[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) g_l'(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X})\{h_{g,l}(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\} \boldsymbol{X}]$. The desired result follows with $\boldsymbol{I}_{\gamma}(\boldsymbol{\psi}, \boldsymbol{\xi}) = \boldsymbol{I}_{\gamma}^{(1)}(\boldsymbol{\psi}, \boldsymbol{\xi}) + \boldsymbol{I}_{\gamma}^{(2)}(\boldsymbol{\psi}, \boldsymbol{\xi})$. □

Proof of Theorem 3.2. Under condition (C5), $\ell_{\beta}^2(\boldsymbol{\psi}, \boldsymbol{\xi})$, $\ell_{\psi}^2(\boldsymbol{\psi})[\boldsymbol{h}_1]$, $\ell_{\xi}^2(\boldsymbol{\xi})[\boldsymbol{h}_2]$, and $\boldsymbol{\ell}_{\gamma}^{*2}(\boldsymbol{\psi})$ are Glivenko-Cantelli classes over the space of $(\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{h}_1, \boldsymbol{h}_2)$. Thus, the variance estimator $\hat{V}$ is consistent if $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{h}}_{\psi}, \hat{\boldsymbol{h}}_{\xi})$ are consistent. The consistency of $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ follows from Lemma 3.1 and condition (C4). A consistent estimator of $\boldsymbol{h}_{\psi}$ can be obtained by solving $\mathbb{P}_n \ell_{\psi\psi}(\hat{\boldsymbol{\psi}})[\boldsymbol{h}_{\psi}, \boldsymbol{\psi}] = \mathbb{P}_n \ell_{\beta\psi}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}})[\boldsymbol{\psi}]$ for all $\boldsymbol{\psi}$. The estimator $\hat{\boldsymbol{h}}_{\xi} \equiv (\hat{h}_{g,1}, \ldots, \hat{h}_{g,L}, \hat{\boldsymbol{h}}_{\eta,1}, \ldots, \hat{\boldsymbol{h}}_{\eta,L})$ is obtained by solving the empirical

version of (3.4) for the B-spline approximation of the function w. Specifically, $\hat{h}_{g,l} = \sum_{k=1}^{K_n} \hat{\theta}_{lk} B_k$ for $l = 1, \ldots, L$, such that $\hat{\boldsymbol{\theta}}_l \equiv (\hat{\theta}_{l1}, \ldots, \hat{\theta}_{lK_n})$ and $\hat{\boldsymbol{h}}_{\eta,l}$ solve

$$\mathbb{P}_n\big[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) B_{lj}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}) \{\hat{h}_{g,l}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}\big] = \mathbb{P}_n\{A(Y, \boldsymbol{X}; \hat{\boldsymbol{\psi}})(1-R) I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) B_j(\hat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X})\}$$
$$\mathbb{P}_n\big[RI(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \tilde{\boldsymbol{Z}} \{\hat{h}_{g,l}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{h}_{\eta,l}^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}\big] = \mathbb{P}_n\{A(Y, \boldsymbol{X}; \hat{\boldsymbol{\psi}})(1-R) I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \tilde{\boldsymbol{Z}}\}$$

for $j = 1, \ldots, K_n$. By the Glivenko-Cantelli properties of the functions involved in the above equations and the consistency of $\hat{\boldsymbol{\psi}}$ and the B-spline approximations, $\hat{\boldsymbol{h}}_\xi \to \boldsymbol{h}_\xi$. The asymptotic distribution of the test statistic follows from Slutsky's theorem.

We show that the resulting projection $\ell_\xi(\boldsymbol{\xi}^*)[\hat{\boldsymbol{h}}_\xi]$ is the same as that obtained by treating the B-spline terms as fixed covariates. Let $\boldsymbol{Z}_l^{(R)}$ and $\boldsymbol{Z}_l^{(1-R)}$ be matrices formed by stacking $(\boldsymbol{B}(\hat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{X}_i)^{\mathrm{T}}, \tilde{\boldsymbol{Z}}_i^{\mathrm{T}})$ together for subjects with $I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l, R_i = 1) = 1$ and $I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l, R_i = 0) = 1$, respectively, and let $\boldsymbol{A}_l^{(1-R)}$ be a vector with elements $A(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\psi}})$ for subjects with $\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l$ and $R_i = 0$. The estimators are

$$(\hat{\boldsymbol{\theta}}_l^{\mathrm{T}}, \hat{\boldsymbol{h}}_{\eta,l}^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{Z}_l^{(R)\mathrm{T}} \boldsymbol{Z}_l^{(R)})^{-1} (\boldsymbol{Z}_l^{(1-R)\mathrm{T}} \boldsymbol{A}^{(1-R)}).$$

Let $\boldsymbol{Z}_{li}(\boldsymbol{\gamma}) = I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l)(\boldsymbol{B}(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}_i)^{\mathrm{T}}, \tilde{\boldsymbol{Z}}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{Z}_l$ be the generic variable. The projection of the score statistic $\ell_\xi(\boldsymbol{\xi})[\hat{\boldsymbol{h}}_\xi]$ is equal to

$$\begin{aligned}
&\Big[S - \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{g_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}\Big] \sum_{l=1}^{L} (\hat{\boldsymbol{\theta}}_l^{\mathrm{T}}, \hat{\boldsymbol{h}}_{\eta,l}^{\mathrm{T}}) \boldsymbol{Z}_l(\boldsymbol{\gamma}_0) \\
= &\Big[S - \sum_{l=1}^{L} I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l)\{g_l(\boldsymbol{\gamma}_0^{\mathrm{T}} \boldsymbol{X}) + \boldsymbol{\eta}_l^{\mathrm{T}} \tilde{\boldsymbol{Z}}\}\Big] \sum_{l=1}^{L} \Big\{ \sum_{i=1}^{n} (1 - R_i) I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l) A(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\psi}}) \boldsymbol{Z}_{li}(\hat{\boldsymbol{\gamma}})^{\mathrm{T}} \Big\} \\
&\qquad \times \Big\{ \sum_{i=1}^{n} R_i I(\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{x}}_l) \boldsymbol{Z}_{li}(\hat{\boldsymbol{\gamma}}) \boldsymbol{Z}_{li}(\hat{\boldsymbol{\gamma}})^{\mathrm{T}} \Big\}^{-1} \boldsymbol{Z}_l(\boldsymbol{\gamma}_0),
\end{aligned}$$

which is the projection of the score of $\boldsymbol{\xi}$, with $\boldsymbol{Z}_{li}$ treated as a set of fixed covariates. □

3.6.2 Explicit Forms of Variance Estimators

In this section, we formulate the variance estimators for the linear model, the logistic model, and the Cox proportional hazards model. Let $\boldsymbol{Z}_i(\boldsymbol{\gamma}) \equiv (\boldsymbol{Z}_{1i}(\boldsymbol{\gamma})^{\mathrm{T}}, \ldots, \boldsymbol{Z}_{Li}(\boldsymbol{\gamma})^{\mathrm{T}})^{\mathrm{T}}$ denote the vector of predictors of S, with $\boldsymbol{Z}_{li}(\boldsymbol{\gamma})$ defined in the proof of Theorem 3.2. Let $\boldsymbol{\xi}$ denote the

corresponding regression parameter and $\hat{\boldsymbol{\xi}}$ denote the least-squares estimator of $\boldsymbol{\xi}$. Let $\boldsymbol{Z}(\boldsymbol{\gamma}) \equiv (\boldsymbol{Z}_1(\boldsymbol{\gamma})^{\mathrm{T}}, \ldots, \boldsymbol{Z}_L(\boldsymbol{\gamma})^{\mathrm{T}})^{\mathrm{T}}$ be the generic version of $\boldsymbol{Z}_i(\boldsymbol{\gamma})$. The robust imputation score statistic is

$$U_\beta^{\mathrm{rob}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}}) = n^{-1/2} \sum_{i=1}^{n} \ell_{\beta,i}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}}) = n^{-1/2} \sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \hat{\boldsymbol{\psi}}) \{R_i S_i + (1 - R_i) \hat{\boldsymbol{\xi}}^{\mathrm{T}} \boldsymbol{Z}_i(\hat{\boldsymbol{\gamma}})\}.$$

Let

$$\begin{aligned}
\boldsymbol{\ell}_{\xi,i}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= R_i \{S_i - \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}_i(\boldsymbol{\gamma})\} \boldsymbol{Z}_i(\boldsymbol{\gamma}) \\
\boldsymbol{I}_{\xi\gamma}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= -\mathrm{E}\{R \boldsymbol{Z}(\boldsymbol{\gamma}) \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}^{(1)}(\boldsymbol{\gamma}) \boldsymbol{X}^{\mathrm{T}}\} \\
\boldsymbol{I}_{\xi\xi}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= -\mathrm{E}\{R \boldsymbol{Z}(\boldsymbol{\gamma}) \boldsymbol{Z}(\boldsymbol{\gamma})^{\mathrm{T}}\} \\
\boldsymbol{I}_{\beta\xi}(\boldsymbol{\psi}, \boldsymbol{\xi}) &= \mathrm{E}\{(1 - R) A(Y, \boldsymbol{X}; \boldsymbol{\psi}) \boldsymbol{Z}(\boldsymbol{\gamma})\},
\end{aligned}$$

where $\boldsymbol{Z}^{(1)}(\boldsymbol{\gamma}) = (\boldsymbol{Z}_1^{(1)}(\boldsymbol{\gamma})^{\mathrm{T}}, \ldots, \boldsymbol{Z}_L^{(1)}(\boldsymbol{\gamma})^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{Z}_l^{(1)}(\boldsymbol{\gamma}) = (I(\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_l) \boldsymbol{B}_l'(\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X})^{\mathrm{T}}, \boldsymbol{0}_q^{\mathrm{T}})^{\mathrm{T}}$ for $l = 1, \ldots L$, the derivative $\boldsymbol{B}'$ is defined component-wise, and $\boldsymbol{0}_q$ is a q-vector of zeros.

First, we consider the linear and logistic models. For the linear model, we redefine $A(Y, \boldsymbol{X}; \boldsymbol{\psi}) = Y - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}$, where the error variance is omitted because it only acts as a scaling factor. For the two models, there is no nuisance parameter $\boldsymbol{\zeta}$, and $\boldsymbol{\psi} = \boldsymbol{\gamma}$. Let

$$\begin{aligned}
\boldsymbol{\ell}_{\gamma,i}^{\mathrm{GLM}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) &= A(Y_i, \boldsymbol{X}_i; \boldsymbol{\gamma}) \boldsymbol{X}_i, \\
\boldsymbol{I}_{\gamma\gamma}^{\mathrm{GLM}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) &= \mathrm{E}\{\boldsymbol{A}^{(1)}(\boldsymbol{X}; \boldsymbol{\gamma}) \boldsymbol{X}^{\mathrm{T}}\} \\
\boldsymbol{I}_{\beta\gamma}^{\mathrm{GLM}}(\boldsymbol{\gamma}, \boldsymbol{\xi}) &= \mathrm{E}[\boldsymbol{A}^{(1)}(\boldsymbol{X}; \boldsymbol{\gamma})^{\mathrm{T}} \{RS + (1 - R) \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}(\boldsymbol{\gamma})\} + (1 - R) A(Y, \boldsymbol{X}; \boldsymbol{\gamma}) \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Z}^{(1)}(\boldsymbol{\gamma}) \boldsymbol{X}^{\mathrm{T}}],
\end{aligned}$$

where $\boldsymbol{A}^{(1)}(\boldsymbol{X}; \boldsymbol{\gamma}) = -\boldsymbol{X}$ for the linear model, and $\boldsymbol{A}^{(1)}(\boldsymbol{X}; \boldsymbol{\gamma}) = -e^{\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}} / (1 + e^{\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}})^2 \boldsymbol{X}$ for the logistic model. In the sequel, we may omit the arguments of the above functions. The variance estimator of $U_\beta^{\mathrm{rob}}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\gamma}})$ is

$$n^{-1} \sum_{i=1}^{n} \{(\ell_{\beta,i} - \bar{\ell}_\beta) - (\hat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{GLM}} - \hat{\boldsymbol{I}}_{\beta\xi}^{\mathrm{T}} \hat{\boldsymbol{I}}_{\xi\xi}^{-1} \hat{\boldsymbol{I}}_{\xi\gamma})(\hat{\boldsymbol{I}}_{\gamma\gamma}^{\mathrm{GLM}})^{-1} (\boldsymbol{\ell}_{\gamma,i}^{\mathrm{GLM}} - \bar{\boldsymbol{\ell}}_\gamma^{\mathrm{GLM}}) - \hat{\boldsymbol{I}}_{\beta\xi}^{\mathrm{T}} \hat{\boldsymbol{I}}_{\xi\xi}^{-1} (\boldsymbol{\ell}_{\xi,i} - \bar{\boldsymbol{\ell}}_\xi)\}^2 \big|_{(\boldsymbol{\gamma},\boldsymbol{\xi}) = (\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\xi}})},$$

where $(\bar{\ell}_\beta, \bar{\boldsymbol{\ell}}_\psi^{\mathrm{GLM}}, \bar{\boldsymbol{\ell}}_\xi)$ is the sample mean of $(\ell_{\beta,i}, \boldsymbol{\ell}_{\psi,i}^{\mathrm{GLM}}, \boldsymbol{\ell}_{\xi,i})$, and $(\hat{\boldsymbol{I}}_{\gamma\gamma}^{\mathrm{GLM}}, \hat{\boldsymbol{I}}_{\xi\gamma}, \hat{\boldsymbol{I}}_{\xi\xi}, \hat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{GLM}}, \hat{\boldsymbol{I}}_{\beta\xi})$ is the empirical version of $(\boldsymbol{I}_{\gamma\gamma}^{\mathrm{GLM}}, \boldsymbol{I}_{\xi\gamma}, \boldsymbol{I}_{\xi\xi}, \boldsymbol{I}_{\beta\gamma}^{\mathrm{GLM}}, \boldsymbol{I}_{\beta\xi})$.

For the Cox proportional hazards model, let

$$\boldsymbol{\mathcal{K}}^{(r)}(t;\boldsymbol{\gamma},\boldsymbol{\xi}) = n^{-1}\sum_{j=1}^{n} I(\tilde{T}_j \geq t) e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_j} \begin{pmatrix} \boldsymbol{X}_j \\ R_j S_j + (1-R_j)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}_j(\boldsymbol{\gamma}) \end{pmatrix}^{\otimes r}$$

for $r = 0$, 1, and 2, where $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for any vector $\boldsymbol{a}$. Partition $\boldsymbol{\mathcal{K}}^{(1)} = (\boldsymbol{\mathcal{K}}_1^{(1)\mathrm{T}}, \mathcal{K}_2^{(1)})^{\mathrm{T}}$ with $\boldsymbol{\mathcal{K}}_1^{(1)} \in \mathbb{R}^{q\times 1}$ and $\mathcal{K}_2^{(1)} \in \mathbb{R}$ and $\boldsymbol{\mathcal{K}}^{(2)}$ into $\boldsymbol{\mathcal{K}}_{11}^{(2)} \in \mathbb{R}^{q\times q}$, $\boldsymbol{\mathcal{K}}_{12}^{(2)} \in \mathbb{R}^{q\times 1}$, $\boldsymbol{\mathcal{K}}_{21}^{(2)} \in \mathbb{R}^{1\times q}$, and $\mathcal{K}_{22}^{(2)} \in \mathbb{R}$. Let

$$\begin{aligned}
\boldsymbol{\ell}_{\gamma,i}^{\mathrm{COX}}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \Delta_i\left\{\boldsymbol{X}_i - \frac{\boldsymbol{\mathcal{K}}_1^{(1)}(\tilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\tilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}\right\} - \frac{1}{n}\sum_{j=1}^{n}\frac{I(\tilde{T}_j \leq \tilde{T}_i)\Delta_j e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_j}}{\mathcal{K}^{(0)}(\tilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\left\{\boldsymbol{X}_i - \frac{\boldsymbol{\mathcal{K}}_1^{(1)}(\tilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\tilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\right\} \\
\ell_{\Lambda,i}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \Delta_i\frac{\mathcal{K}_2^{(1)}(\tilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\tilde{T}_i;\boldsymbol{\gamma},\boldsymbol{\xi})} - \Lambda(\tilde{T}_i)e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_i}\{R_iS_i + (1-R_i)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}_i(\boldsymbol{\gamma})\} \\
&\quad + \frac{1}{n}\sum_{j=1}^{n}\frac{I(\tilde{T}_j \leq \tilde{T}_i)\Delta_j e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_j}}{\mathcal{K}^{(0)}(\tilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\left\{R_iS_i + (1-R_i)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}_i(\boldsymbol{\gamma}) - \frac{\mathcal{K}_2^{(1)}(\tilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}{\mathcal{K}^{(0)}(\tilde{T}_j;\boldsymbol{\gamma},\boldsymbol{\xi})}\right\} \\
\boldsymbol{I}_{\beta\gamma}^{\mathrm{COX}}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \mathrm{E}[\{\Delta - \Lambda(\tilde{T})e^{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}}\}(1-R)\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{Z}^{(1)}(\boldsymbol{\gamma})\boldsymbol{X}^{\mathrm{T}}] + \boldsymbol{v}_{12}(\boldsymbol{\gamma},\boldsymbol{\xi})^{\mathrm{T}} \\
\boldsymbol{I}_{\gamma\gamma}^{\mathrm{COX}}(\boldsymbol{\psi},\boldsymbol{\xi}) &= \boldsymbol{v}_{11}(\boldsymbol{\gamma},\boldsymbol{\xi}),
\end{aligned}$$

where

$$\boldsymbol{v}_{jk}(\boldsymbol{\gamma};\boldsymbol{\xi}) = -\mathrm{E}\left[\Delta\left\{\frac{\boldsymbol{\mathcal{K}}_{jk}^{(2)}(\tilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})^{\mathrm{T}}}{\mathcal{K}^{(0)}(\tilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})} - \frac{\boldsymbol{\mathcal{K}}_j^{(1)}(\tilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})\boldsymbol{\mathcal{K}}_k^{(1)}(\tilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})^{\mathrm{T}}}{\mathcal{K}^{(0)}(\tilde{T};\boldsymbol{\gamma},\boldsymbol{\xi})^2}\right\}\right] \quad \text{for } j,\ k = 1 \text{ or } 2.$$

The variance estimator of $U_\beta^{\mathrm{rob}}(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\gamma}})$ is

$$n^{-1}\sum_{i=1}^{n}\{(\ell_{\beta,i} - \bar{\ell}_\beta) - (\hat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{COX}} - \hat{\boldsymbol{I}}_{\beta\xi}\hat{\boldsymbol{I}}_{\xi\xi}^{-1}\hat{\boldsymbol{I}}_{\xi\gamma})(\hat{\boldsymbol{I}}_{\gamma\gamma}^{\mathrm{COX}})^{-1}(\boldsymbol{\ell}_{\gamma,i}^{\mathrm{COX}} - \bar{\boldsymbol{\ell}}_\gamma^{\mathrm{COX}}) - (\ell_{\Lambda,i} - \bar{\ell}_\Lambda) - \hat{\boldsymbol{I}}_{\beta\xi}\hat{\boldsymbol{I}}_{\xi\xi}^{-1}(\boldsymbol{\ell}_{\xi,i} - \bar{\boldsymbol{\ell}}_\xi)\}^2,$$

with $(\boldsymbol{\psi},\boldsymbol{\xi})$ evaluated at $(\hat{\boldsymbol{\psi}},\hat{\boldsymbol{\xi}})$, where $(\bar{\boldsymbol{\ell}}_\gamma^{\mathrm{COX}}, \bar{\ell}_\Lambda^{\mathrm{COX}})$ is the sample mean of $(\boldsymbol{\ell}_{\gamma,i}^{\mathrm{COX}}, \ell_{\Lambda,i}^{\mathrm{COX}})$, and $(\hat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{COX}}\hat{\boldsymbol{I}}_{\gamma\gamma}^{\mathrm{COX}})$ is the empirical version of $(\boldsymbol{I}_{\beta\gamma}^{\mathrm{COX}}\boldsymbol{I}_{\gamma\gamma}^{\mathrm{COX}})$.

3.6.3 Bias of the Standard Variance Estimator

In this section, we evaluate the standard model-based variance estimator based on imputed data. Consider a generalized linear model with no nuisance parameter. The information matrix of $(\beta,\boldsymbol{\gamma})$ under $\beta = 0$ is

$$\begin{pmatrix} B(\boldsymbol{X};\boldsymbol{\gamma})S^2 & B(\boldsymbol{X};\boldsymbol{\gamma})S\boldsymbol{X}^{\mathrm{T}} \\ B(\boldsymbol{X};\boldsymbol{\gamma})S\boldsymbol{X} & B(\boldsymbol{X};\boldsymbol{\gamma})\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \end{pmatrix},$$

where $B(\boldsymbol{X};\boldsymbol{\gamma}) = \mathrm{Var}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}) \mid \boldsymbol{X}\}$. The limit of the model-based variance estimator with the imputed data is equal to

$$\begin{aligned} \hat{V}^{\mathrm{std}} &= \mathrm{E}[B(\boldsymbol{X};\boldsymbol{\gamma}_0)\{\tilde{S}^2 - \mathrm{E}(\tilde{S}\boldsymbol{X})^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\tilde{S}\boldsymbol{X}\}] \\ &= \mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\tilde{S}^2\} - \mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)S\boldsymbol{X}^{\mathrm{T}}\}\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}^{-1}\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)S\boldsymbol{X}\}, \end{aligned}$$

where $\tilde{S}$ is equal to S if $R = 1$ and is equal to $\tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)$ otherwise.

We derive the bias of the model-based variance estimator under a correct imputation model, i.e., $\mathrm{E}(S \mid \boldsymbol{X}) = \mathrm{E}\{\tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*) \mid \boldsymbol{X}\}$, and a balanced sampling scheme of S, i.e., $\mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)R \mid \boldsymbol{X}\} = 0$. Assume that $\sup_{\boldsymbol{Z}} |\tilde{S}(\boldsymbol{Z};\hat{\boldsymbol{\xi}}) - \tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)| = O_p(n^{-1/2})$. In this case, the imputation score statistic is

$$\begin{aligned} & U_\beta^{\mathrm{imp}}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\xi}}) \\ = \;& n^{-1/2}\sum_{i=1}^n A(Y_i,\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}})\{R_iS_i + (1-R_i)\tilde{S}(\boldsymbol{Z}_i;\hat{\boldsymbol{\xi}})\} \\ = \;& n^{-1/2}\sum_{i=1}^n A(Y_i,\boldsymbol{X}_i;\boldsymbol{\gamma}_0)\tilde{S}_i - B(\boldsymbol{X}_i;\boldsymbol{\gamma}_0)\tilde{S}_i\boldsymbol{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \\ & \quad + A(Y_i,\boldsymbol{X}_i;\boldsymbol{\gamma}_0)(1-R_i)\{\tilde{S}(\boldsymbol{Z};\hat{\boldsymbol{\xi}}) - \tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)\} + o_p(1) \\ = \;& n^{-1/2}\sum_{i=1}^n A(Y_i,\boldsymbol{X}_i;\boldsymbol{\gamma}_0)[\tilde{S}_i - \mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\tilde{S}\boldsymbol{X}^{\mathrm{T}}\}\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}^{-1}\boldsymbol{X}_i] + o_p(1), \end{aligned}$$

where the third equality follows from the rate of convergence of $\tilde{S}(\boldsymbol{Z};\hat{\boldsymbol{\xi}})$ and the balanced sampling scheme. Therefore, the asymptotic variance of the imputation score statistic is

$$\mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2\tilde{S}^2\} - \mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)S\boldsymbol{X}^{\mathrm{T}}\}\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}^{-1}\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)S\boldsymbol{X}\}.$$

The bias of the model-based estimator is

$$\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)\tilde{S}^2\} - \mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2\tilde{S}^2\}. \tag{3.5}$$

Let $v_{Y,j}(\boldsymbol{X}) = \mathrm{Var}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}) \mid R = j,\boldsymbol{X}\}$, $p_R(\boldsymbol{X}) = P(R = 1 \mid \boldsymbol{X})$, $v_S(\boldsymbol{X}) = \mathrm{E}(S^2 \mid \boldsymbol{X})$, and

$v_{\tilde{S}}(\boldsymbol{X}) = \mathrm{E}\{\tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)^2 \mid \boldsymbol{X}\}$. The first term of (3.5) is $\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)p_R(\boldsymbol{X})v_S(\boldsymbol{X})\} + \mathrm{E}[B(\boldsymbol{X};\boldsymbol{\gamma}_0)\{1 - p_R(\boldsymbol{X})\}v_{\tilde{S}}(\boldsymbol{X})]$. By the definition of $B(\boldsymbol{X};\boldsymbol{\gamma}_0)$,

$$\begin{aligned} B(\boldsymbol{X};\boldsymbol{\gamma}_0) &= \mathrm{Var}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0) \mid \boldsymbol{X}\} \\ &= \mathrm{E}[\mathrm{Var}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0) \mid \boldsymbol{X}, R\} \mid \boldsymbol{X}] + \mathrm{Var}[\mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0) \mid \boldsymbol{X}, R\} \mid \boldsymbol{X}] \\ &= p_R(\boldsymbol{X})v_{Y,1}(\boldsymbol{X}) + \{1 - p_R(\boldsymbol{X})\}v_{Y,0}(\boldsymbol{X}). \end{aligned}$$

The second term of (3.5) is

$$\begin{aligned} & \mathrm{E}[A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2\{RS + (1-R)\tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)\}^2] \\ = \; & \mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2 RS^2\} + \mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2(1-R)\tilde{S}(\boldsymbol{Z};\boldsymbol{\xi}^*)^2\} \\ = \; & \mathrm{E}[v_{Y,1}(\boldsymbol{X})p_R(\boldsymbol{X})v_S(\boldsymbol{X})\} + \mathrm{E}\{v_{Y,0}(\boldsymbol{X})\{1 - p_R(\boldsymbol{X})\}v_{\tilde{S}}(\boldsymbol{X})]. \end{aligned}$$

To see that (3.5) is non-zero in general, consider the simple case of $X = 1$, $v_{Y,1} > v_{Y,0}$, and $v_S > v_{\tilde{S}}$, such that there is no covariate, the variance of the phenotype is larger among subjects with observed S, and the variance of the true S is larger than that of the imputed S. By Chebyshev's sum inequality, the true variance $\mathrm{E}\{A(Y,\boldsymbol{X};\boldsymbol{\gamma}_0)^2\tilde{S}^2\}$ is strictly larger than the limit of the model-based variance estimator $\mathrm{E}\{B(\boldsymbol{X};\boldsymbol{\gamma}_0)^2\tilde{S}^2\}$. By contrast, if the missing-data mechanism does not depend on Y, then $v_{Y,1}$ and $v_{Y,0}$ are equal. As a result, (3.5) is equal to zero, and the model-based variance estimator is consistent.

3.6.4 Evaluation of Power

In this section, we evaluate the power of the imputation score test under the linear model: $Y = \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X} + \beta S + \mathrm{N}(0,\sigma^2)$. Assume the same imputation model for S as in Chapter 3.6.2 with a general predictor $\boldsymbol{Z}(\boldsymbol{\gamma})$. Assume also that $\boldsymbol{X}$ is contained in $\boldsymbol{Z}(\boldsymbol{\gamma})$ and that $\mathrm{E}(S \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \mathrm{E}\{\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) \mid \boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}, \tilde{\boldsymbol{X}}\}$, such that the score statistic is unbiased. The missing-data status can be expressed as $R = I\{(Y,\tilde{\boldsymbol{X}}) \in \Omega(\omega)\}$, where for every fixed ω, $\Omega(\omega)$ is a deterministic subset of the space of $(Y,\tilde{\boldsymbol{X}})$, and ω is a random variable that is independent of $(Y,\boldsymbol{Z}(\boldsymbol{\gamma}),S)$. Adopting the notation introduced in Chapter 3.6.2, the score statistic can be expanded as $n^{-1/2}\sum_{i=1}^n \boldsymbol{\Psi}(\ell_{\beta,i}, \boldsymbol{\ell}_{\gamma,i}^{\mathrm{T}}, \boldsymbol{\ell}_{\xi,i}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\Psi} = (1, (\boldsymbol{I}_{\beta\gamma} - \boldsymbol{I}_{\beta\xi}^{\mathrm{T}}\boldsymbol{I}_{\xi\xi}^{-1}\boldsymbol{I}_{\xi\gamma})\boldsymbol{I}_{\gamma\gamma}^{-1}, -\boldsymbol{I}_{\beta\xi}\boldsymbol{I}_{\xi\xi}^{-1})$, $\boldsymbol{\ell}_{\gamma,i} = \boldsymbol{\ell}_{\gamma,i}^{\mathrm{GLM}}$, $\boldsymbol{I}_{\gamma\gamma} = \boldsymbol{I}_{\gamma\gamma}^{\mathrm{GLM}}$, and $\boldsymbol{I}_{\beta\gamma} = \boldsymbol{I}_{\beta\gamma}^{\mathrm{GLM}}$.

Under the contiguous alternative of $\beta_n = n^{-1/2}b$ for some fixed b,

$$
\begin{aligned}
n^{-1/2}\mathrm{E}\Big\{\sum_{i=1}^{n}\ell_{\beta,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\Big\} &= n^{1/2}\mathrm{E}[(Y-\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X})\{RS+(1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\
&= n^{1/2}\mathrm{E}[\varepsilon R\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]+b\mathrm{E}[S\{RS+(1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}],
\end{aligned}
$$

where $\varepsilon = Y-\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}-\beta_n S$. The first expectation on the far right side of the above expression is

$$
\begin{aligned}
& \mathrm{E}[\varepsilon I\{(Y,\tilde{\boldsymbol{X}})\in\Omega(\omega)\}\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\
= \;& \mathrm{E}[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}P\{(Y,\tilde{\boldsymbol{X}})\in\Omega(\omega)\mid\varepsilon,S,\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\
= \;& \mathrm{E}[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}P\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\tilde{\boldsymbol{X}})\in\Omega(\omega)\mid\varepsilon,\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\tilde{\boldsymbol{X}}\}] \\
& +\beta_n\mathrm{E}\Big[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}S\frac{\partial}{\partial t}P\{(t+\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\tilde{\boldsymbol{X}})\in\Omega(\omega)\mid\varepsilon,\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\tilde{\boldsymbol{X}}\}|_{t=0}\Big]+o(1).
\end{aligned}
$$

The first term of the right side of the last equality is zero by assumption. Thus,

$$
\begin{aligned}
& n^{-1/2}\mathrm{E}\Big\{\sum_{i=1}^{n}\ell_{\beta,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*;\boldsymbol{\gamma}_0)\Big\} \\
\to \;& b\mathrm{E}\Big[\varepsilon\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}S\frac{\partial}{\partial t}P\{(t+\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\tilde{\boldsymbol{X}})\in\Omega(\omega)\mid\varepsilon,\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\tilde{\boldsymbol{X}}\}|_{t=0}\Big] \\
& \quad +b\mathrm{E}[S\{RS+(1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}] \\
\equiv \;& b\mu_\beta.
\end{aligned}
$$

Simple algebraic manipulation yields $n^{-1/2}\mathrm{E}\{\sum_{i=1}^{n}\boldsymbol{\ell}_{\gamma,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*)\}=b\mathrm{E}(S\boldsymbol{X})\equiv b\boldsymbol{\mu}_\gamma$. In addition,

$$
\begin{aligned}
& n^{-1/2}\mathrm{E}\Big\{\sum_{i=1}^{n}\boldsymbol{\ell}_{\xi,i}(\boldsymbol{\gamma}_0,\boldsymbol{\xi}^*)\Big\} \\
= \;& n^{1/2}\mathrm{E}[I\{(Y,\tilde{\boldsymbol{X}})\in\Omega(\omega)\}\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{Z}(\boldsymbol{\gamma}_0)] \\
= \;& n^{1/2}\mathrm{E}[P\{(Y,\tilde{\boldsymbol{X}})\in\Omega(\omega)\mid S,\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{Z}(\boldsymbol{\gamma}_0)] \\
\to \;& b\mathrm{E}\Big[\frac{\partial}{\partial t}P\{(t+\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X}+\varepsilon,\tilde{\boldsymbol{X}})\in\Omega(\omega)\mid\boldsymbol{\gamma}_0^{\mathrm{T}}\boldsymbol{X},\tilde{\boldsymbol{X}}\}|_{t=0}S\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\Big]\equiv b\boldsymbol{\mu}_\xi.
\end{aligned}
$$

Therefore, the asymptotic distribution of the score test statistic is non-central chi-square with non-centrality parameter

$$
\mathcal{C}=\frac{b^2\{\boldsymbol{\Psi}(\mu_\beta,\boldsymbol{\mu}_\gamma^{\mathrm{T}},\boldsymbol{\mu}_\xi^{\mathrm{T}})^{\mathrm{T}}\}^2}{\boldsymbol{\Psi}\mathrm{E}\{(\ell_{\beta,i},\boldsymbol{\ell}_{\gamma,i}^{\mathrm{T}},\boldsymbol{\ell}_{\xi,i}^{\mathrm{T}})^{\otimes 2}\}\boldsymbol{\Psi}^{\mathrm{T}}}.
$$

Because the conditional distribution of S given $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$ may be misspecified, $\{S-\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}$ in

$\boldsymbol{\ell}_{\xi,i}(\boldsymbol{\gamma}_0, \boldsymbol{\xi}^*)$ is generally dependent of $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$. Thus, the non-centrality parameter is a function of high moments of $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$, and it is difficult to evaluate the power for different choices of $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$. Consider the case that S is missing completely at random. In this case, $\boldsymbol{I}_{\beta\gamma} = -\mathrm{E}[\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{X}^{\mathrm{T}}]$, and $\boldsymbol{I}_{\beta\xi} = \boldsymbol{0}$. The denominator of $\mathcal{C}$ is

$$
\begin{aligned}
& \mathrm{Var}(\ell_{\beta,i} - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\ell}_{\gamma,i}) \\
= \; & \mathrm{E}[\varepsilon^2\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{X}\}^2] \\
= \; & \sigma^2\mathrm{E}\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) - \mathrm{E}[\{RS + (1-R)(\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{X}]\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\}^2 \\
= \; & \sigma^2\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]^2,
\end{aligned}
$$

where $\mathcal{P}_{\boldsymbol{X}}^{\perp}$ denotes the projection onto the orthogonal space of $\boldsymbol{X}$, i.e., $\mathcal{P}_{\boldsymbol{X}}^{\perp}T = T - \boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1} \times \mathrm{E}(\boldsymbol{X}T)$ for any random variable T. The numerator of $\mathcal{C}$ is

$$
\begin{aligned}
& b^2(\mu_\beta - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\mu}_\gamma)^2 \\
= \; & b^2(\mathrm{E}\{RS^2 + (1-R)S\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\} - \mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\mathrm{E}[\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}\boldsymbol{X}])^2 \\
= \; & b^2(\mathrm{E}[R\{S^2 - \mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}S\} \\
& \quad + (1-R)\{S\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0) - \mathrm{E}(S\boldsymbol{X}^{\mathrm{T}})\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}])^2 \\
= \; & b^2[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2 + (1-p_R)\mathrm{Cov}\{\mathcal{P}_{\boldsymbol{X}}^{\perp}S, \mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]^2 \\
= \; & b^2[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2 + (1-p_R)\mathrm{E}\{\mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}^2]^2,
\end{aligned}
$$

where $p_R = P(R=1)$, and the last equality follows from the definition of $\boldsymbol{\xi}^*$ and the fact that $\boldsymbol{X}$ is contained in $\boldsymbol{Z}(\boldsymbol{\gamma}_0)$. As a result,

$$
\mathcal{C} = \frac{b^2}{\sigma^2} \times \frac{[p_R\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp}S)^2 + (1-p_R)\mathrm{E}\{\mathcal{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}^2]^2}{\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS + (1-R)\boldsymbol{\xi}^{*\mathrm{T}}\boldsymbol{Z}(\boldsymbol{\gamma}_0)\}]^2}.
$$

To show that the proposed test is more powerful than the imputation score test without stratification or the B-spline terms, we consider two sets of linear predictors $\boldsymbol{Z}_1(\hat{\boldsymbol{\gamma}})$ and $\boldsymbol{Z}_2(\hat{\boldsymbol{\gamma}})$, where $\boldsymbol{Z}_1(\boldsymbol{\gamma})$ is contained in $\boldsymbol{Z}_2(\boldsymbol{\gamma})$. Let $\tilde{S}_1$ and $\tilde{S}_2$ be the imputed values of S using $\boldsymbol{Z}_1(\boldsymbol{\gamma}_0)$ and $\boldsymbol{Z}_2(\boldsymbol{\gamma}_0)$, respectively. The score test with $\boldsymbol{Z}_2(\hat{\boldsymbol{\gamma}})$ in the imputation model is asymptotically more

powerful if and only if

$$\frac{[p_R \mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} S)^2 + (1-p_R)\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} \tilde{S}_2)^2]^2}{\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS + (1-R)\tilde{S}_2\}]^2} \geq \frac{[p_R \mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} S)^2 + (1-p_R)\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} \tilde{S}_1)^2]^2}{\mathrm{E}[\mathcal{P}_{\boldsymbol{X}}^{\perp}\{RS + (1-R)\tilde{S}_1\}]^2}. \tag{3.6}$$

After some algebraic manipulation, the denominator on the left side of (3.6) can be expressed as

$$p_R \mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} S)^2 + (1-p_R)\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} \tilde{S}_2)^2 + p_R(1-p_R)\{\mathrm{E}(S\boldsymbol{X}^{\mathrm{T}}) - \mathrm{E}(\tilde{S}_2\boldsymbol{X}^{\mathrm{T}})\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\{\mathrm{E}(S\boldsymbol{X}) - \mathrm{E}(\tilde{S}_2\boldsymbol{X})\}.$$

Compared to $\tilde{S}_1$, $\tilde{S}_2$ is the projection of S onto a larger linear space. Thus,

$$\begin{aligned} & \{\mathrm{E}(S\boldsymbol{X}^{\mathrm{T}}) - \mathrm{E}(\tilde{S}_2\boldsymbol{X}^{\mathrm{T}})\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\{\mathrm{E}(S\boldsymbol{X}) - \mathrm{E}(\tilde{S}_2\boldsymbol{X})\} \\ \leq \; & \{\mathrm{E}(S\boldsymbol{X}^{\mathrm{T}}) - \mathrm{E}(\tilde{S}_1\boldsymbol{X}^{\mathrm{T}})\}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\{\mathrm{E}(S\boldsymbol{X}) - \mathrm{E}(\tilde{S}_1\boldsymbol{X})\}, \end{aligned}$$

and $\mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} \tilde{S}_2)^2 \geq \mathrm{E}(\mathcal{P}_{\boldsymbol{X}}^{\perp} \tilde{S}_1)^2$. It follows that (3.6) holds, and the test with a larger set of covariates in the imputation is more powerful.

CHAPTER 4

SURVIVAL TIME PREDICTION WITH MULTI-PLATFORM HIGH-DIMENSIONAL GENOMIC DATA

4.1 Introduction

Prediction of disease outcomes, such as individual patient survival time, is critically important for cancer patients. Traditional prognostic models that rely solely on clinical variables, such as age and tumor stage, fail to account for the molecular features of tumors and thus may lead to suboptimal treatment decisions (Shedden et al. 2008). To remedy this situation, many studies have incorporated gene expression data for survival prediction (West et al. 2001; Beer et al. 2002; Shipp et al. 2002; van't Veer et al. 2002).

Large-scale genomics projects, such as The Cancer Genome Atlas (TCGA), have generated detailed molecular data on patients with a variety of cancer types. In TCGA, six different types of genomic data have been collected on patients: DNA copy number variation, DNA somatic mutation, mRNA expression, miRNA expression, DNA methylation, and the expression of ∼200 proteins/phosphoproteins. The availability of different data types has enabled researchers to address a variety of important questions. For example, patients can be more precisely classified into molecular subtypes based on integrative clustering of multiple genomic data types or platforms (Shen et al. 2009; Mo et al. 2013; Lock et al. 2013). In addition, it is possible to identify genes that are related to patient survival time by decomposing the expression of each gene into a component that is explained by the methylation level and a component that is not (Wang et al. 2013).

One unsolved issue in cancer genomics is the prognostic value of integrated genomic and clinical data versus clinical data only. Yuan et al. (2014) compared models with clinical data only versus models with both clinical and genomic data on various cancer types and concluded that genomic data provide only a limited gain in survival prediction accuracy. In their analysis, however, potential differences among data types were not taken into account. For breast cancer, for instance, the combination of genomic and clinical data *does* improve outcome predictions (Parker et al. 2009; Fan et al. 2011). One of the goals of the present work is to more fully explore the predictive power of

integrating clinical and genomic data together.

The second issue that we wish to address is the prognostic value of individual gene expression data (∼12,000) versus a predefined set of gene expression signatures, which are also referred to as "modules" (∼500). Gene modules represent activated molecular signaling pathways or specific biological processes, and they have been developed to better capture signaling pathway activity or cell type heterogeneity within tumors. We wish to investigate whether individual gene expression data or existing gene modules provide more accurate outcome prediction.

A third issue is the relative importance of different types of genomic data in outcome prediction. Different data types are collected at different costs and also with widely varying feature spaces. Naturally, not all data types are equally important in outcome prediction. We aim to determine which data types may be omitted from analysis without a significant reduction in prediction accuracy.

The overarching methodological challenge that we address in this chapter is the identification of genomic variables predictive of survival time or other clinical outcomes when the number of variables is much larger than the sample size. Penalization methods, such as least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) and elastic net (Zou and Hastie 2005), are commonly used to identify important genomic variables. Elastic net is preferable to LASSO because it better handles highly correlated variables (Zou and Hastie 2005). However, both LASSO and elastic net are generic variable selection procedures that do not distinguish different types of data. It is highly desirable to account for differences in predictive power and the number of features of different data types.

An alternative to penalization-based variable selection methods is boosting (Bühlmann 2006; Bühlmann and Yu 2006). Boosting originated from a machine-learning context for classification problems (Freund and Schapire 1996), which is not a separate classification procedure but is a method to improve existing procedures. Specifically, boosting combines many "weak" classifiers to obtain a much "stronger" classifier by repeatedly fitting the weak classifier and reweighting the samples based on the results of the previous fit. Over the years, boosting has been generalized to a much broader statistical framework that accommodates continuous, binary, and censored outcomes; see Hastie et al. (2009) for the development of boosting. The main advantage of boosting is that it can incorporate general simple estimation procedures, such as component-wise least-square regression and classification tree, with stable computation. Boosting is typically performed with

repeating a classification/prediction procedure on the same set of data, and boosting with different data types treated separately in a variable-selection context has not been explored.

In this chapter, we develop a novel method, termed Integrative Boosting (I-Boost), that combines elastic net with statistical boosting. The I-Boost approach considers each data type separately, so that small but predictive data types will not be dominated by larger ones. Herein, we evaluate I-Boost using simulation studies and data from the TCGA on patients with eight different cancer types, namely colon adenocarcinoma (COAD), rectal adenocarcinoma (READ), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), kidney renal cell carcinoma (KIRC), and head and neck squamous cell carcinoma (HNSC). We show that I-Boost outperforms LASSO and elastic net in terms of survival prediction. We also demonstrate that when I-Boost is applied to integrated genomic and clinical datasets, the accuracy of prediction is improved in comparison to that achieved with the use of a single data type.

In Chapter 4.2, we review LASSO and elastic net and present the I-Boost procedure. In Chapter 4.3, we report results from simulation studies that compare the proposed and existing penalization methods. In Chapter 4.4, we present results from the analyses of the TCGA dataset and address the aforementioned three issues. We make some concluding remarks in Chapter 4.5 and provide details about the TCGA data in Chapter 4.6.

4.2 Methods

4.2.1 LASSO and Elastic Net

Both LASSO (Tibshirani 1996) and elastic net (Zou and Hastie 2005) are penalized estimation procedures; LASSO is a special case of elastic net. The elastic net estimator for the Cox proportional hazards model maximizes the objective function

$$\log L(\boldsymbol{\beta}) - \lambda\Big\{\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1}{2}(1-\alpha)\sum_{j=1}^{p} \beta_j^2\Big\},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of regression parameters, L is the partial likelihood, λ and α are tuning parameters, and p is the number of covariates. The objective function for LASSO is the above expression with $\alpha = 1$. With $\alpha = 0$, elastic net reduces to ridge regression. For large values

of λ, the penalty term dominates, and the parameter estimates tend to be small with some values being exactly zero. Unlike LASSO, elastic net exhibits the grouping effect in that the regression parameters for a group of highly correlated variables tend to be equal, which is desirable in the context of gene selection.

The tuning parameters α and λ are typically selected by K-fold cross-validation. Cross-validation was suggested by the authors of LASSO and elastic net (Tibshirani 1996; Zou and Hastie 2005). In a K-fold cross-validation procedure, the dataset is randomly split into K subsets of equal size. Each training set is a combination of $(K-1)$ subsets, and the corresponding testing set is the remaining subset. For each set of tuning parameters, LASSO or elastic net is performed on all training sets, and the resulting estimates are evaluated on the corresponding testing sets. We set the cross-validation error associated with each set of tuning parameters to be the average deviance based on the partial likelihood over all testing sets. The set of tuning parameters with the least cross-validation error is chosen for the analysis on the whole dataset.

4.2.2 I-Boost

Let T be the survival time of interest, C be the censoring time, $Y \equiv \min(T, C)$ be the observed survival or censoring time, and $\Delta \equiv I(T \leq C)$ be the event indicator, where $I(\cdot)$ is the indicator function. Let $\boldsymbol{X} \equiv (\boldsymbol{X}^{(1)\mathrm{T}}, \ldots, \boldsymbol{X}^{(K)\mathrm{T}})^{\mathrm{T}}$ be the set of predictors, where $\boldsymbol{X}^{(k)}$ is a p_k-vector that contains predictors for the kth data type for $k = 1, \ldots, K$, and K is the total number of data types. The observed data consist of $(Y_i, \Delta_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$.

The I-Boost algorithm is given as follows:

1. Set $f_{0,i} = 0$ for $i = 1, \ldots, n$, and let $\boldsymbol{f}_0 = (f_{0,1}, \ldots, f_{0,n})^{\mathrm{T}}$.

2. Consider $m = 1, 2, \ldots$:

   (a) For $k = 1, \ldots, K$, calculate

$$\boldsymbol{\beta}^{(k)} = \mathrm{argmax}_{\boldsymbol{\beta}} \{ \log L_n^{(k)}(\boldsymbol{f}_{m-1}; \boldsymbol{\beta}) - p^{(k)}(\boldsymbol{\beta}; \alpha, \lambda) \}$$

   using the coordinate-descent method (Simon et al. 2011), where

$$L_n^{(k)}(\boldsymbol{f}; \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} \left(\frac{e^{f_i + \boldsymbol{X}_i^{(k)\mathrm{T}} \boldsymbol{\beta}}}{\sum_{j:Y_j \geq Y_i} e^{f_j + \boldsymbol{X}_j^{(k)\mathrm{T}} \boldsymbol{\beta}}} \right)^{\Delta_i}$$

is the partial likelihood with offset term $\boldsymbol{f}$ and covariates $\boldsymbol{X}^{(k)}$, $\boldsymbol{f} = (f_1, \ldots, f_n)^{\mathrm{T}}$, and $p^{(k)}(\boldsymbol{\beta}; \alpha, \lambda) \equiv \lambda\{\alpha \sum_{j=1}^{p_k} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{p_k} \beta_j^2\}$ is the elastic net penalty.

(b) Let $k^* = \mathrm{argmax}_k \log L_n^{(k)}(\boldsymbol{f}_{m-1}; \boldsymbol{\beta}^{(k)})$.

(c) Set $f_{m,i} = f_{m-1,i} + v\boldsymbol{X}_i^{(k^*)\mathrm{T}}\boldsymbol{\beta}^{(k^*)}$ for $i = 1, \ldots, n$ and some fixed $0 < v \leq 1$ and $\boldsymbol{f}_m = (f_{m,0}, \ldots, f_{m,n})^{\mathrm{T}}$.

At the kth iteration, the current estimate $v\boldsymbol{\beta}^{(k^*)}$ contributes to the final parameter estimate additively. The final parameter estimate for each data type is simply the sum of the current estimates obtained across all steps in which the data type achieves the maximum at step 2(b).

We propose two versions of I-Boost, with different methods for selecting the tuning parameters α and λ at step 2(a). For I-Boost-CV, we adopt a two-dimensional five-fold cross-validation separately at each iteration. To keep the update at each iteration small, we restrict the search of λ on a set of large values. This results in small estimates and also a large number of parameters being shrunk to zero. We set $v = 1$ at step 2(c).

For I-Boost-Permutation, we adopt the permutation method proposed by Sabourin et al. (2015). The procedure is motivated by the principle that in a null model, i.e., in the absence of any relevant predictors, the tuning parameters should be chosen such that no variable is selected. The permutation selection procedure generates hypothetical null models by randomly permuting $(Y_i, \Delta_i, f_{m-1,i})$ B times, so that in each permutation dataset the association between the predictors and the outcome (and the offset term) is removed. For each permutation, we fix $\alpha = 1$ and find the smallest λ such that no variable is selected. The selected λ for that iteration is the median of the B values of λ. We set $v = 0.1$ at step 2(c).

Conventional boosting methods require a stopping criterion to avoid over-fitting. In our experience, however, because the tuning parameters are selected separately at each iteration for the two procedures, they eventually lead to shrinkage of all (current) parameter estimates. Therefore, we do not adopt a separate procedure to determine the stopping time of the iteration. We terminate the iteration when $\boldsymbol{f}_m$ remains constant for five consecutive iterations.

4.3 Simulation Studies

We used the R-package "sampling" (Tillé and Matei 2016) to sample 500 subjects who had complete data from the TCGA pan-cancer data set, balancing the clinical variables between the

sampled and non-sampled subjects; see Chapter 4.6 for a description of the data. The outcome variable was generated from a proportional hazards model with clinical variables, gene modules, copy number data, miRNA expression, protein expression, and somatic mutation data as predictors. The censoring time was generated to be independent of the predictors and survival time, with a censoring proportion of approximately 65%.

The regression parameters were chosen to result in different proportions of signals across data types, where the signal of data type k is defined to be $\mathrm{Var}(\boldsymbol{X}^{(k)\mathrm{T}}\boldsymbol{\beta}_0^{(k)})$, $\boldsymbol{\beta}_0^{(k)}$ is the true regression parameter value, and the predictors were standardized. The variables with non-zero regression parameters, hereafter referred to as signal variables, were chosen to be weakly correlated. We considered three settings, with the distributions of signals and number of signal variables shown in Figure 4.1. In Setting 1, the clinical variables contain much stronger signals than the other data types. The miRNA and protein variables do not contain any signal. In Setting 2, the clinical variables contain the most signals, and the remaining signals are evenly distributed across the other data types. In Setting 3, the clinical variables contain the most signals, the modules and copy number variables contain the second largest amount of signals, and the protein variables do not contain any signals. In all three settings, the number of signal variables is less than 4% of the total number of variables.

We evaluated the performance of LASSO, elastic net, and the two versions of I-Boost. Five-fold cross validation was used to select the tuning parameters for LASSO and elastic net. For elastic net, the cross-validation was performed over a two-dimensional grid of (α, λ), while for LASSO, we set $\alpha = 1$. The grid for α for elastic net was chosen to be $(0.05, 0.1, 0.2, \ldots, 1.0)$, and a grid for λ was chosen separately for each α. To make the selection procedure more stable, we repeated the split and evaluation procedure five times, and the cross-validation errors were averaged over the five repetitions.

We assessed the performance of the methods by the quality of variable selection and prediction. For variable selection, we report the false discovery rate. Because elastic net tends to select variables that are highly correlated, we considered the selection of a variable that is highly correlated with any of the signal variables (with absolute correlation greater than or equal to 0.6) as a "true discovery", so as not to bias against elastic net. For each selected variable, we calculated the maximum of the absolute correlation between the selected variable and the signal variables, and we report the mean

of the maximum absolute correlation over all selected variables. We call this measure the mean correlation.

For prediction, we report the correlation between the estimated risk score $\sum_{k=1}^{K} \boldsymbol{X}^{(k)\mathrm{T}} \hat{\boldsymbol{\beta}}^{(k)}$ and the true risk score $\sum_{k=1}^{K} \boldsymbol{X}^{(k)\mathrm{T}} \boldsymbol{\beta}_0^{(k)}$, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the estimated parameter vector. A higher correlation represents a larger degree of agreement between the predicted and actual outcomes. We call this assessment method the risk correlation.

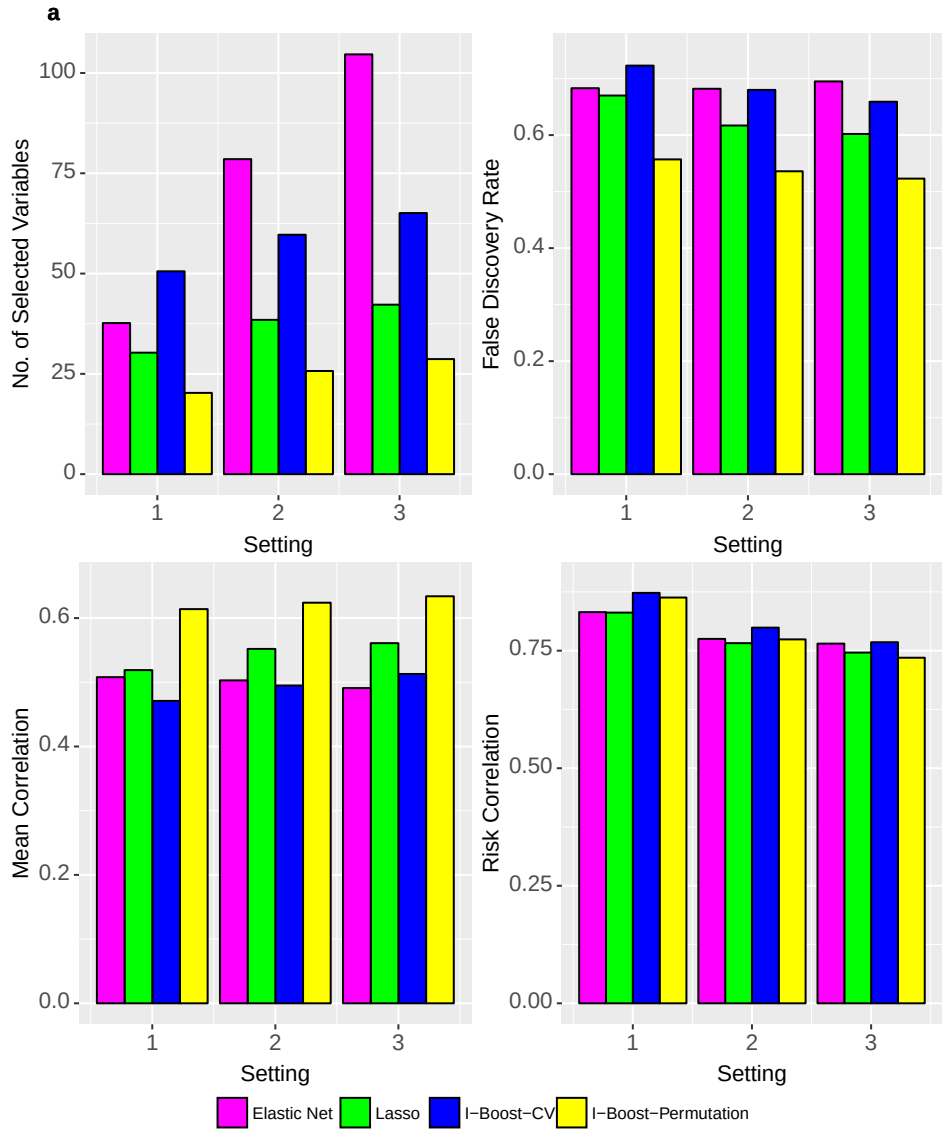
The performance of elastic net, LASSO, and the two versions of I-Boost based on 1,000 replications is shown in Figure 4.1. We also present the average number of variables selected for each method. The simulation results show that, on average, elastic net selects the largest number of variables, followed by I-Boost-CV, LASSO, and I-Boost-Permutation. In all settings, I-Boost-Permutation has the lowest false discovery rate and selects variables that are, on average, most strongly correlated with the signal variables.

For prediction, the two I-Boost methods perform the best overall. In Settings 1 and 2, where the clinical variables contribute a large proportion of signals, the I-Boost methods produce more accurate prediction than both elastic net and LASSO. In Setting 3, I-Boost-CV performs similarly to elastic net, while LASSO performs worse than both versions of I-Boost. Between the two versions of I-Boost, I-Boost-CV tends to yield better prediction than I-Boost-Permutation, possibly because of the larger number of variables selected by I-Boost-CV. Thus, if the main interest is the selection of relevant variables, then one might consider I-Boost-Permutation for more conservative variable selection, even though this method is somewhat inferior in prediction when compared to I-Boost-CV.

4.4 Data Analysis Results

4.4.1 Evaluation of LASSO, Elastic Net, and I-Boost Using TCGA Data

We evaluated the performance of the analysis procedures using three TCGA datasets, namely, the LUAD dataset, the KIRC dataset, and the pan-cancer dataset derived from more than 1,400 patients with one of eight different tumor types; see Chapter 4.6 for a description of the data. To assess an analysis procedure, we split the data into multiple training and testing sets with a 3:2 ratio of sample sizes. We used the R-package "sampling" (Tillé and Matei 2016) to balance the data split on the clinical variables. We performed the analysis on the training sets, and the results were assessed on the corresponding testing sets. In particular, we estimated the C-index (Pencina and



b

	Clinical	Modules	CNV	Mutation	miRNA	Protein
No. of Variables	4	20	10	5	5	5
Setting 1	70%	20%	5%	5%	0%	0%
Setting 2	50%	10%	10%	10%	10%	10%
Setting 3	40%	20%	20%	10%	10%	0%

Figure 4.1. Simulation Settings and Results: (a) Performance of LASSO, Elastic Net, I-Boost-CV, and I-Boost-Permutation, in Terms of Number of Variables Selected, False Discovery Rate, Mean Correlation, and Risk Correlation Under Three Different Settings; and (b) Number of Signal Variables and Distribution of Signals Across Different Data Types for the Three Simulation Settings. The number of signal variables is zero if the proportion of signals of the data type is 0%. Abbreviations are as follows: GeneExp represents raw gene expression; Module represents gene module; Clinical represents clinical variable; CNV represents copy number variant; Mutation represents somatic mutation; miRNA represents micro-RNA expression; and Protein represents protein expression.

D'Agostino 2004) on the testing set with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the parameter estimate obtained from the training set. In the case of no variable being selected, a C-index value of 0.5 was assigned. For each split of the data, we repeated this estimation-validation procedure on different combinations of data types as predictors. From the seven available data types, we formed 95 unique combinations of data types. (Raw gene expression and gene modules did not enter the same model.) Finally, the analyses were conducted on the multiple splits of the data and on the 95 combinations of data types for the LUAD, KIRC, and pan-cancer datasets.

The average C-index values over the splits obtained from LASSO and elastic net are given in Figures 4.2–4. For the analyses of the pan-cancer and KIRC datasets, the prediction tends to be much better than random guessing, i.e., the C-index values are much larger than 0.5. In the analyses of the LUAD dataset, which has smaller sample size, some of the C-index values are close to or only slightly larger than 0.5.

For many models, the predictive performance of elastic net is either similar or superior to LASSO. The tuning parameter α was selected to be smaller than 0.5 for over 70% of the time using the cross-validation procedure. For approximately 15% of the time, $\alpha = 1$ was selected.

For LASSO and elastic net, the models containing more data types as predictors do not necessarily perform better than those with fewer data types. One possible explanation is that the extra data types may contain very little relevant information on patient survival, such that adding those data types introduces more noise into the model than signal. In practice, however, it is challenging to decide which data types to consider without prior knowledge of their importance.

Figures 4.5–7 show the average values of the C-index obtained by I-Boost-CV and elastic net for different models. For the LUAD and pan-cancer datasets, I-Boost provides better prediction in many cases, especially for models that include clinical variables. This finding is consistent with the conclusions from the simulation studies. For the KIRC dataset, the difference is not as clear.

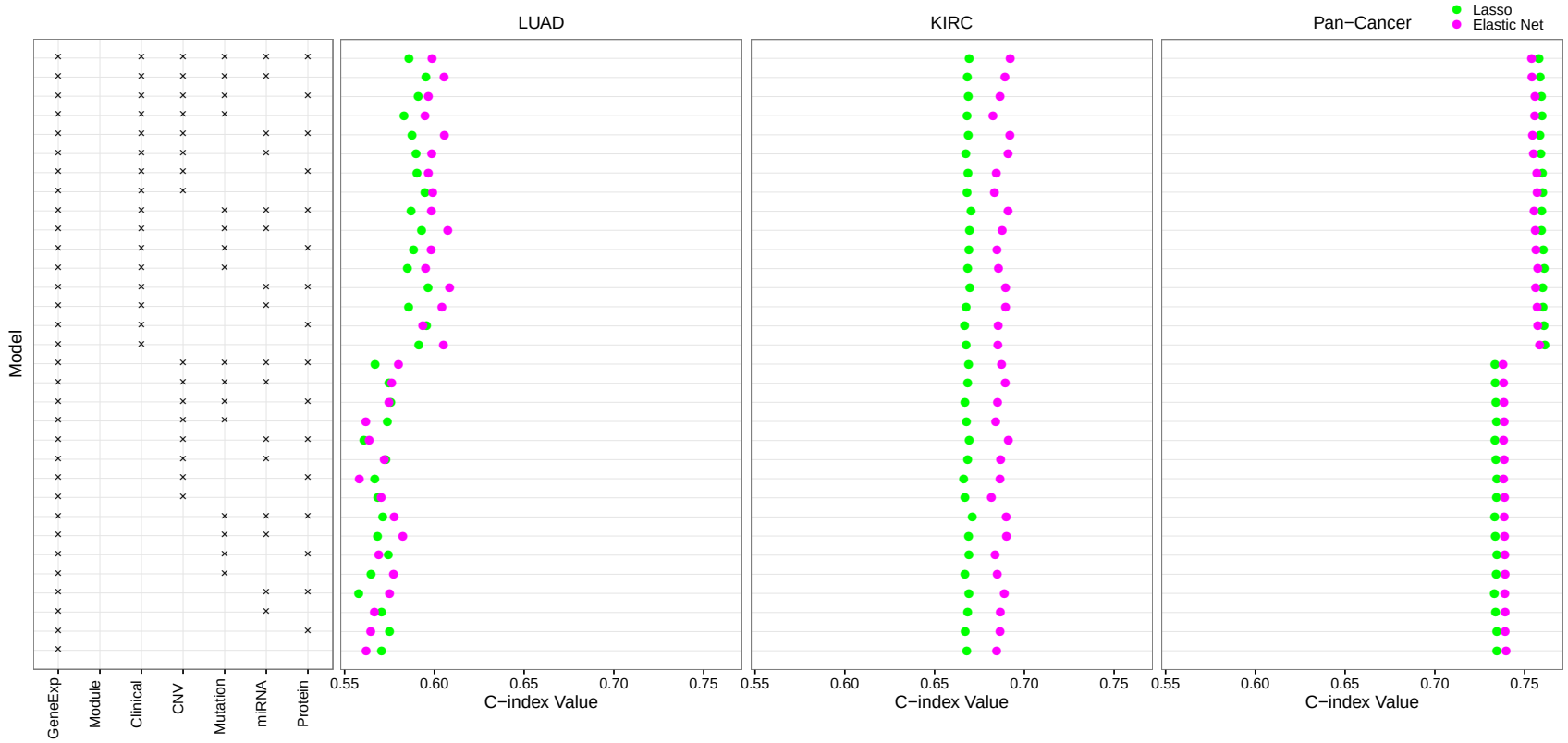


Figure 4.2. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using LASSO and Elastic Net for Models With Raw Gene Expression Data. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing LASSO or elastic net on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

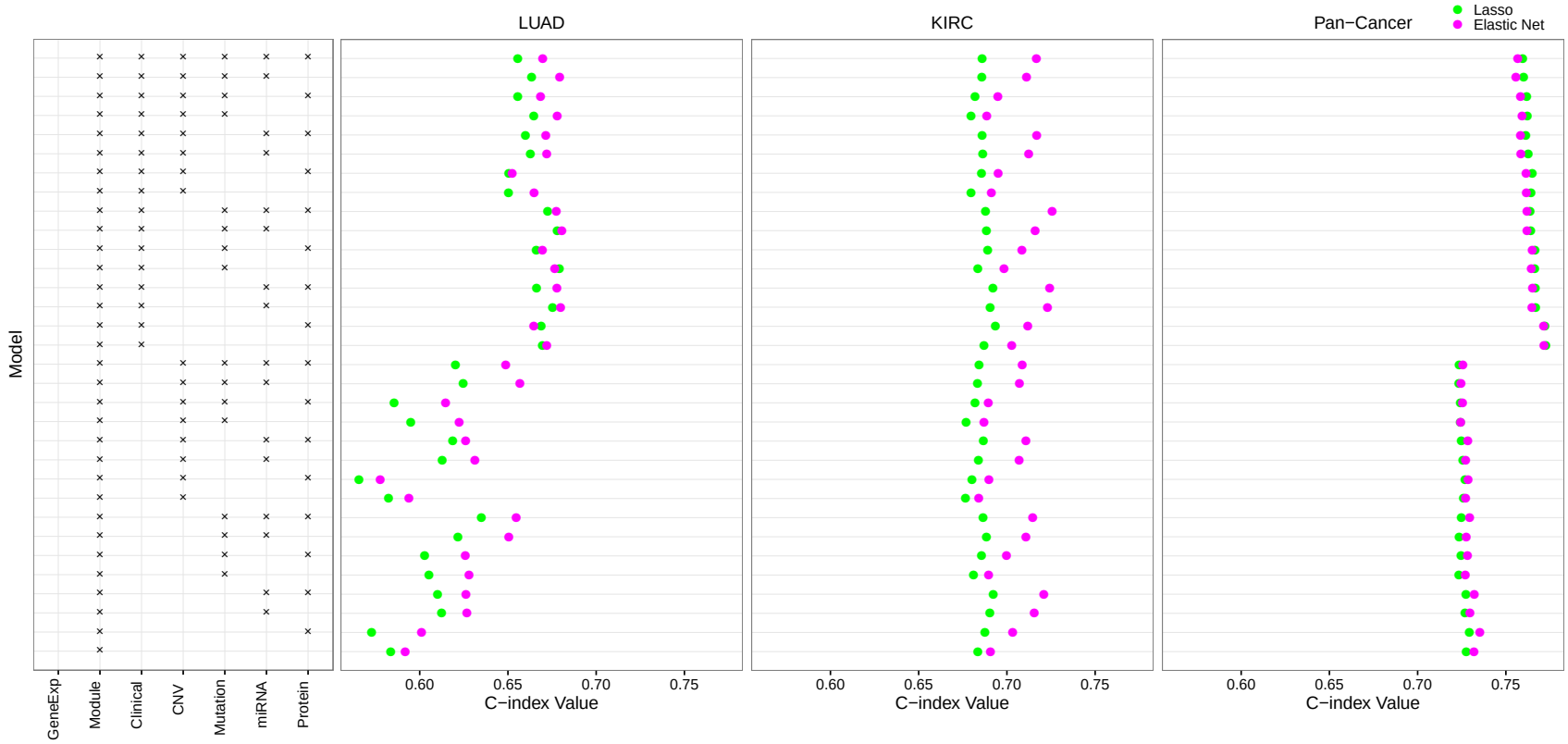


Figure 4.3. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using LASSO and Elastic Net for Models With Gene Modules. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing LASSO or elastic net on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

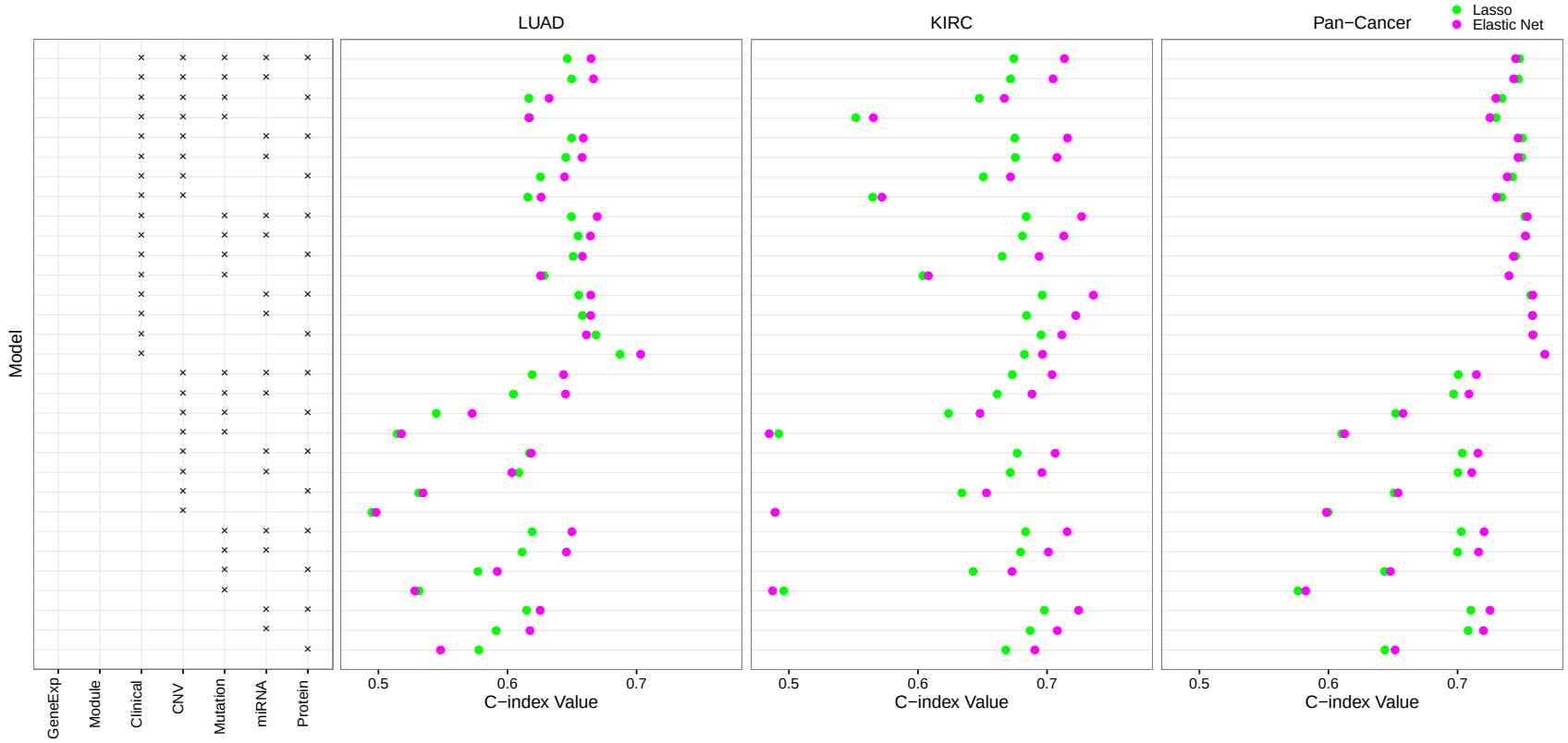


Figure 4.4. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using LASSO and Elastic Net for Models Without Gene Expression Data. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing LASSO or elastic net on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

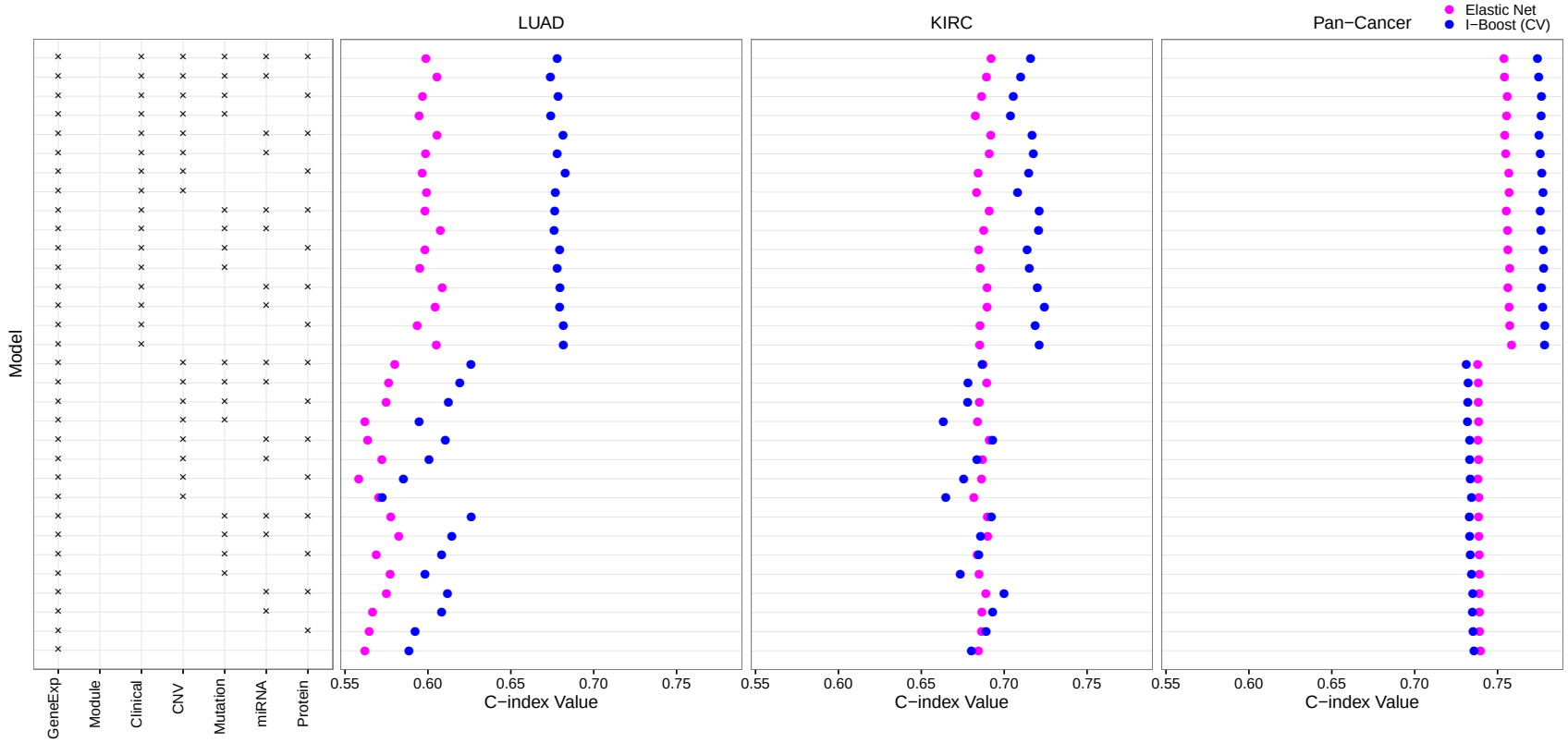


Figure 4.5. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets Using Elastic Net and I-Boost-CV for Models With Raw Gene Expression Data. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing elastic net or I-Boost-CV on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

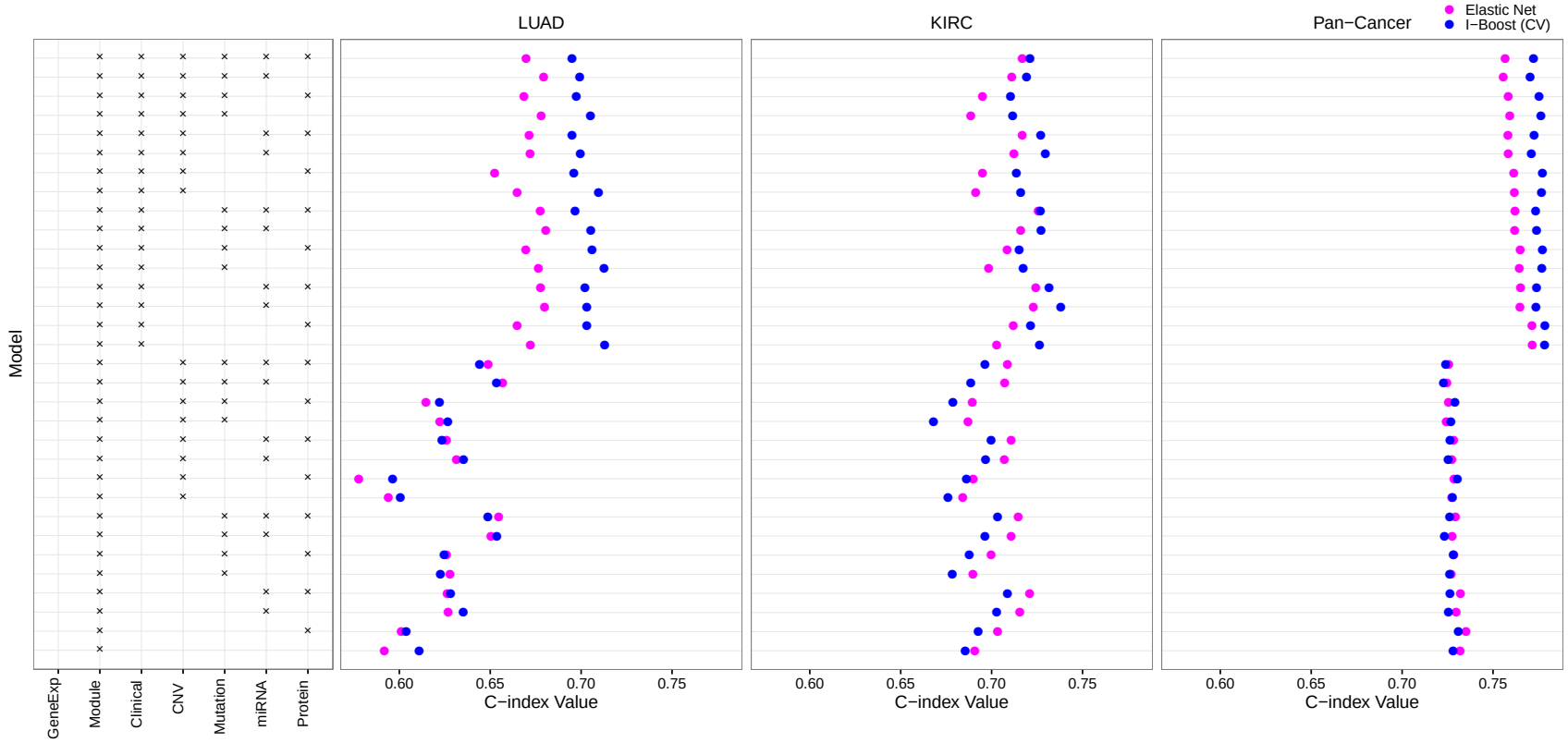


Figure 4.6. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets Using Elastic Net and I-Boost-CV for Models With Gene Modules. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing elastic net or I-Boost-CV on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

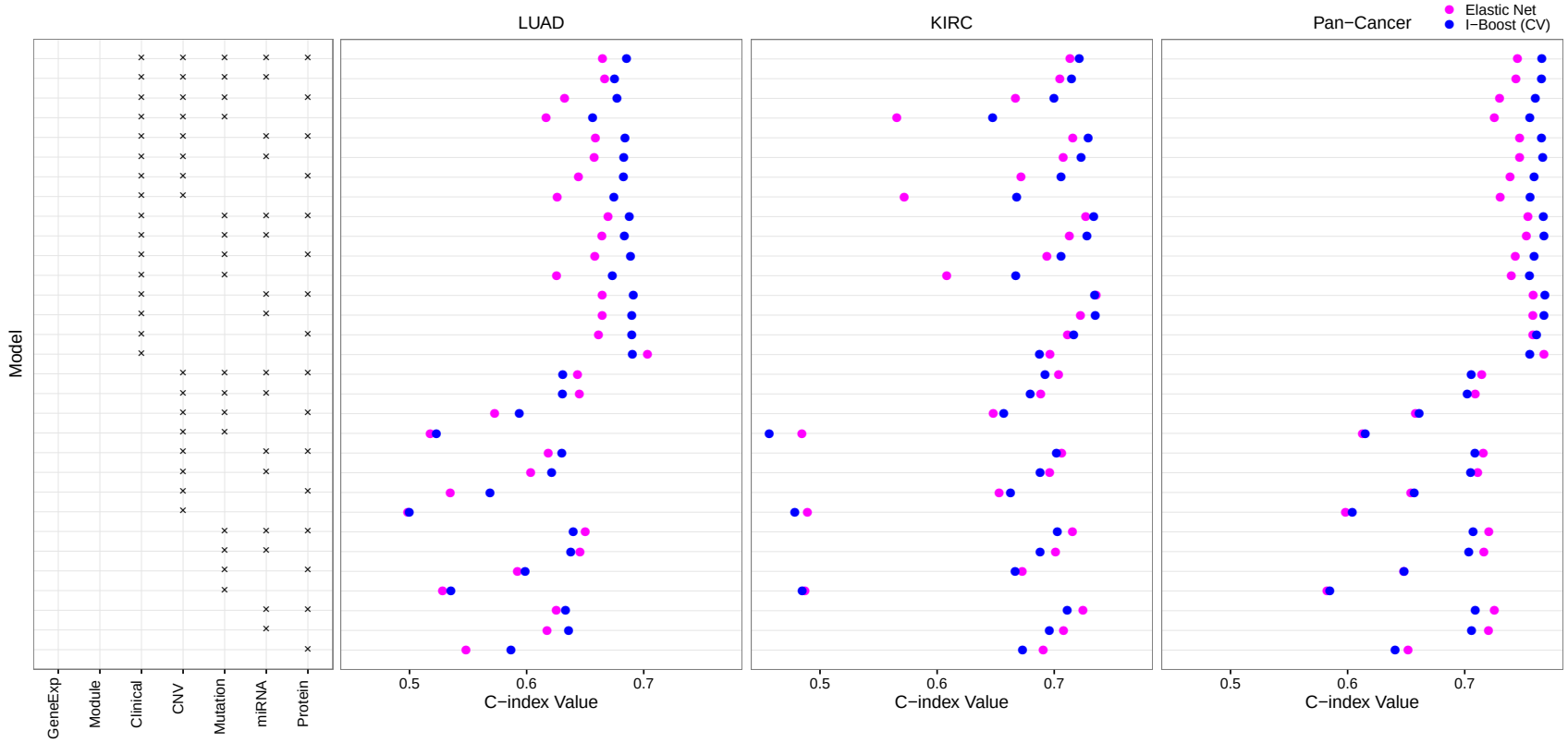


Figure 4.7. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets Using Elastic Net and I-Boost-CV for Models Without Gene Expression Data. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing elastic net or I-Boost-CV on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

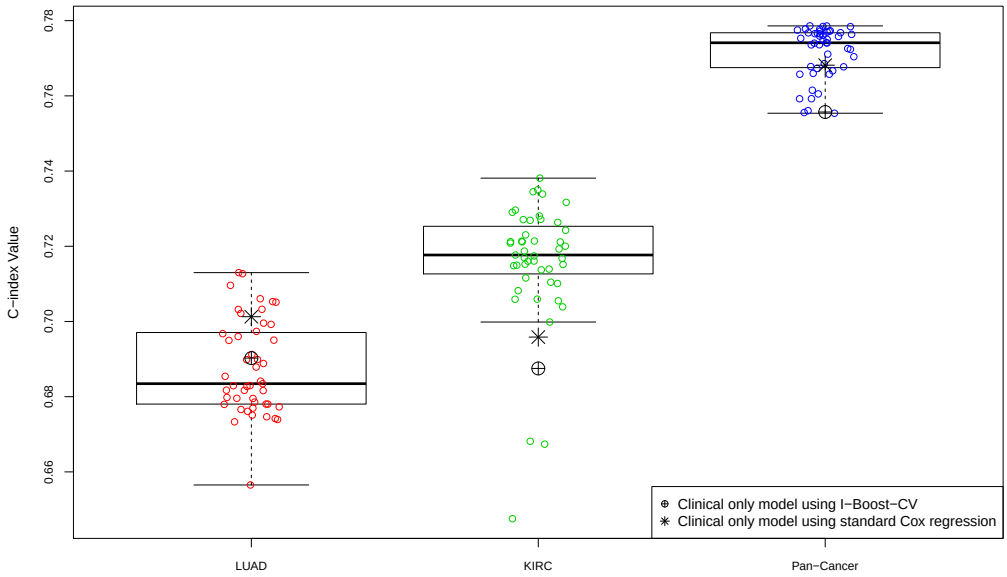


Figure 4.8. C-Index Values for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using I-Boost-CV for Models Containing Clinical Variables. Each dot represents the average C-index value obtained by performing I-Boost-CV on a set of predictors that contains the clinical variables over 30 training and testing dataset pairs. The average C-index values obtained by fitting I-Boost-CV or the standard Cox regression on the clinical variables are marked.

To assess whether I-Boost captures useful information from the genomic variables beyond that drawn from the clinical variables, we compared the values of the C-index obtained by I-Boost-CV for all models that include the clinical variables. The plot of the C-index values is provided in Figure 4.8. Because the standard maximum partial likelihood estimation is arguably preferable to any regularized regression procedures in the model with clinical variables only, we also computed the C-index from the standard analysis for that model. For the KIRC and pan-cancer datasets, the majority of the models that contain both clinical and genomic variables provide better prediction than either I-Boost or maximum partial likelihood estimation with clinical variables only. For the LUAD dataset, only several models that contain both clinical and genomic variables provide better prediction than the model with clinical variables only. These results indicate that genomic variables contribute to survival prediction in the presence of clinical variables, and the magnitude of the contribution can be large. However, when the same comparisons are made using LASSO or elastic net, the inclusion of genomic variables in the models does not appreciably improve prediction.

4.4.2 Evaluation of Signatures, Individual Genes, and Different Genomic Data Types

To compare the performance of gene modules versus individual gene expression data, we calculated the C-index values for models with each data type separately. Specifically, for each combination of data types besides individual gene expression data and gene modules, we computed the difference between the C-index values obtained from I-Boost-CV on those data types with gene modules and on those data types with gene expression data. The differences in the C-index are shown in Figure 4.9. For the LUAD dataset, the use of gene modules mostly leads to better prediction than the use of gene expression data of all individual genes. For the KIRC and pan-cancer datasets, performance is similar with gene modules or individual gene expression data included in the models.

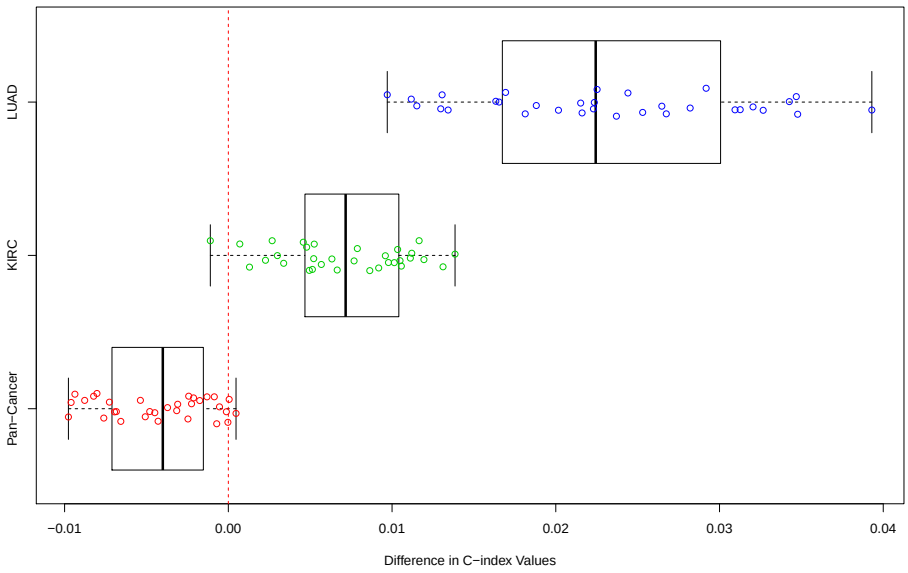


Figure 4.9. Comparison of C-Index Values for Models Containing Raw Gene Expression Data and Models Containing Gene Modules Using the TCGA LUAD, KIRC, and Pan-Cancer Datasets. Each dot represents the difference in average C-index values obtained by fitting I-Boost-CV on two sets of predictors over 30 training and testing dataset pairs. The first set of predictors contains a combination of data types and gene modules; the second set of predictors contains the same combination of data types and raw gene expression data. A positive difference represents better prediction using the model with gene modules.

To evaluate the relative importance of the data types, we constructed a series of nested models as follows. First, we compared all models containing a single data type and selected the model that was the most predictive. Then, we compared the models containing the selected data type and another data type and again selected the most predictive model. This process was repeated until all

data types were included, and the model selected at each step contained all of the previously selected data types. Individual gene expression data were not considered in this analysis. The order that the data types entered the models reflects their relative importance. We performed this procedure for elastic net and the two versions of I-Boost. For the LUAD, KIRC, and pan-cancer datasets, the C-index values for the series of models are plotted in Figure 4.10, and the data type selected at each step is shown. We also plotted the average number of variables selected for each model.

For the LUAD, KIRC, and pan-cancer datasets, with the inclusion of each new data type under I-Boost-Permutation, the C-index tends to increase or stay approximately the same. I-Boost-CV yields results with similar patterns, although the C-index may decrease by a small amount as more data types are included. This indicates that I-Boost extracts useful information from each additional data type and that its performance tends not to be worsened by the inclusion of additional variables. For the LUAD and pan-cancer datasets, clinical variables and gene modules are always the first data types to be selected, and the improvement in prediction accuracy with the inclusion of additional data types is marginal. For the KIRC dataset, miRNA expression data are the first to be selected by elastic net and I-Boost-CV, while gene modules are first selected by I-Boost-Permutation. For elastic net, there is no clear improvement in prediction accuracy when more data types are included. In the LUAD, KIRC, and pan-cancer datasets, elastic net yields larger C-index values for models with a single data type than I-Boost. This result is not surprising, because the main advantage of I-Boost lies in the handling of multiple data types.

I-Boost-Permutation always selects the smallest number of variables, and I-Boost-CV selects the second smallest number of variables in most cases. As more data types are included, the number of variables selected by elastic net tends to fluctuate greatly, while the number of variables selected by I-Boost tends to increase gradually. This suggests that the tuning parameter selection procedure for I-Boost is more stable than that for elastic net. Because the C-index obtained by I-Boost is higher in most cases than that obtained by elastic net, we conclude that I-Boost provides the same or better prediction using fewer variables than elastic net.

To obtain a final set of important predictors, we performed I-Boost-Permutation on the LUAD, KIRC, and pan-cancer datasets. We plotted the comparison of C-index values in Figures 4.11–13 and demonstrate that I-Boost-Permutation yields comparable prediction accuracy to I-Boost-CV. The final models are shown in Tables 4.1–3 for the LUAD, KIRC, and pan-cancer datasets, respectively.

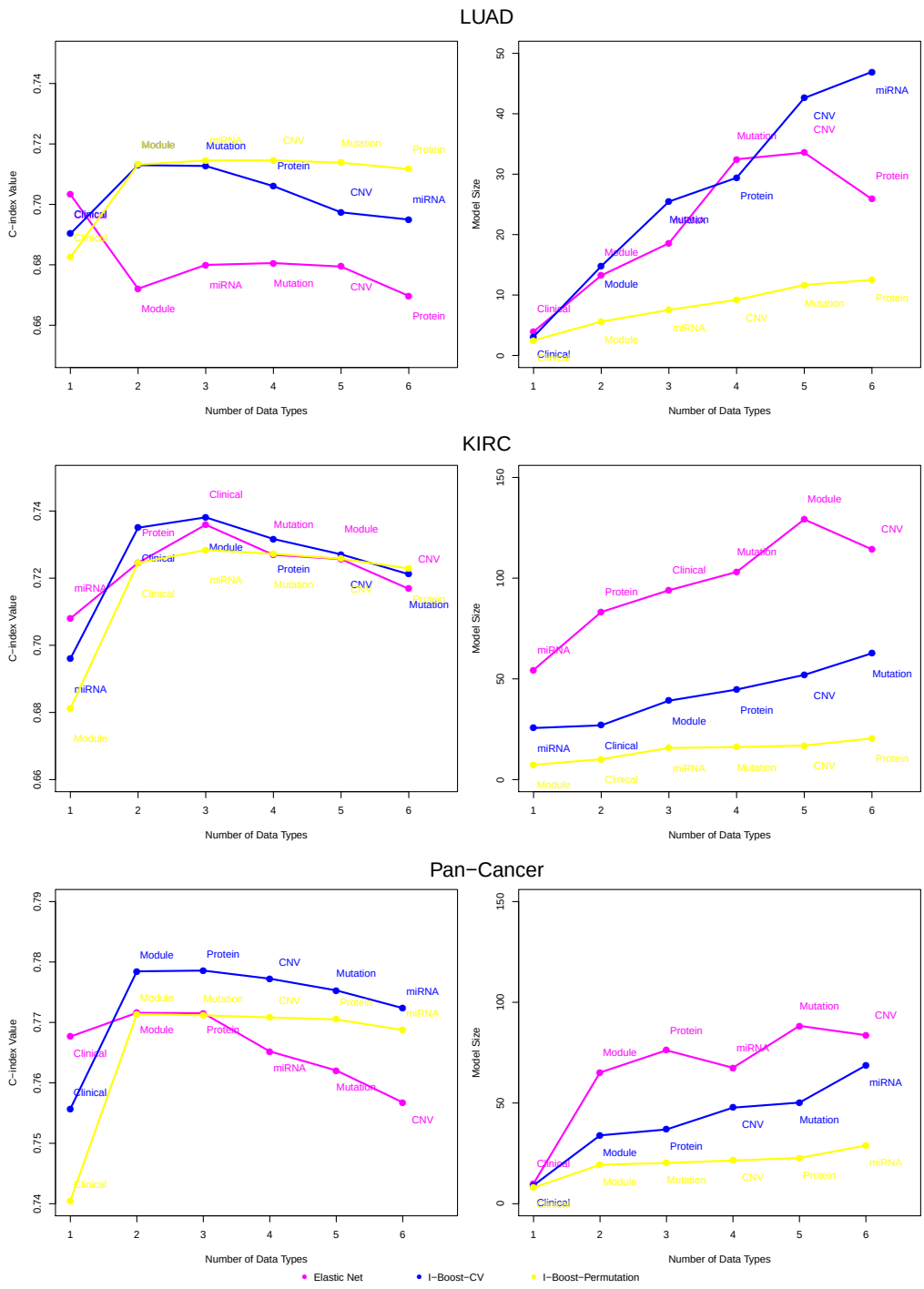


Figure 4.10. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using Elastic Net, I-Boost-CV, and I-Boost-Permutation on Nested Models. For the plots on the left side, each dot represents the average C-index value obtained by fitting elastic net or I-Boost-CV over 30 training and testing dataset pairs. The leftmost dot represents the largest average C-index value among models that contain one data type. Each of the other dots represents the largest average C-index value among models that contain one more data type than the model corresponding to the dot on the left. For the plots on the right side, the average number of selected variables for the models shown on the left is plotted. For all plots, beside each dot, the name of the additional data type is included. See the caption of Figure 4.1 for the abbreviations of the data types.

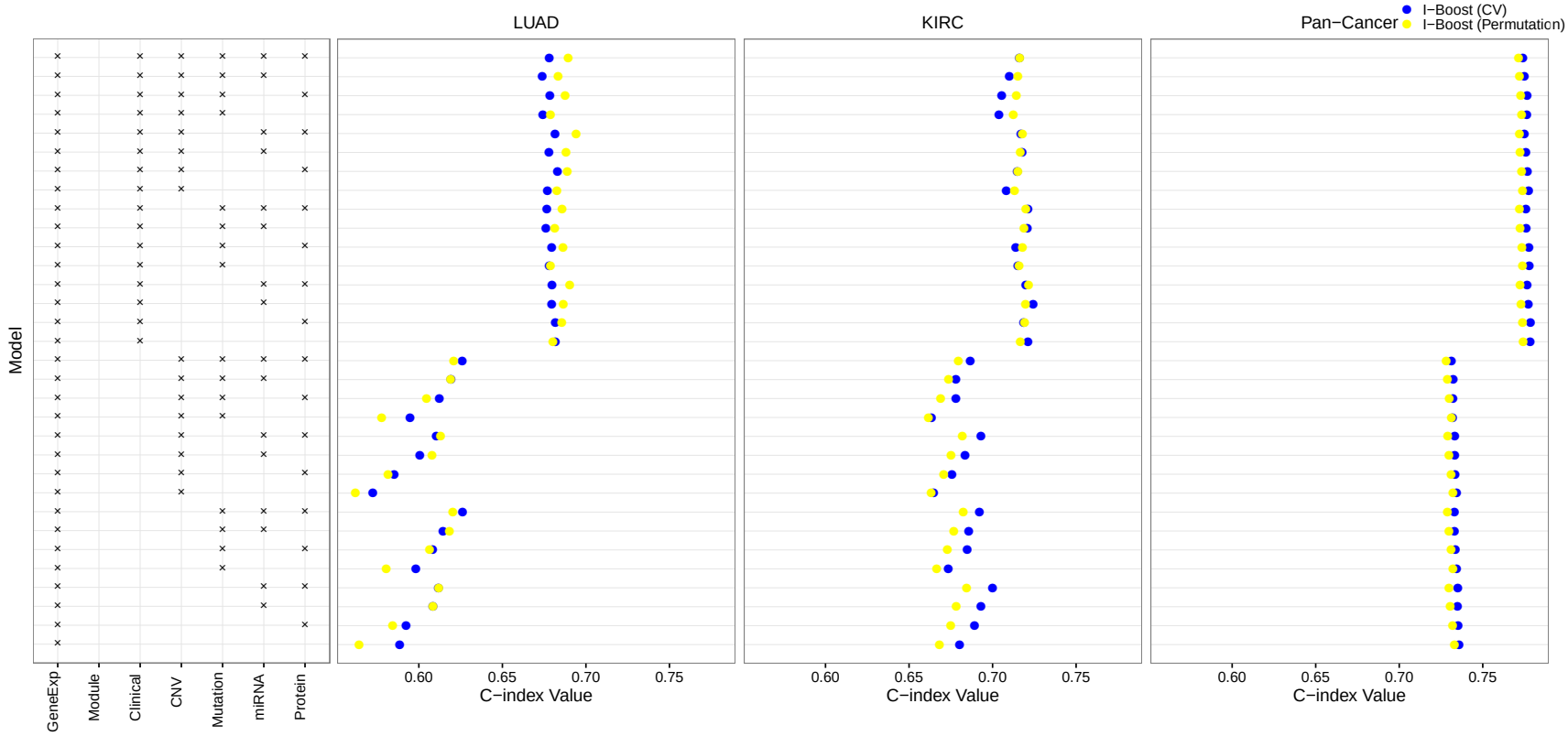


Figure 4.11. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using I-Boost-CV and I-Boost-Permutation for Models With Raw Gene Expression Data. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing I-Boost-CV or I-Boost-Permutation on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

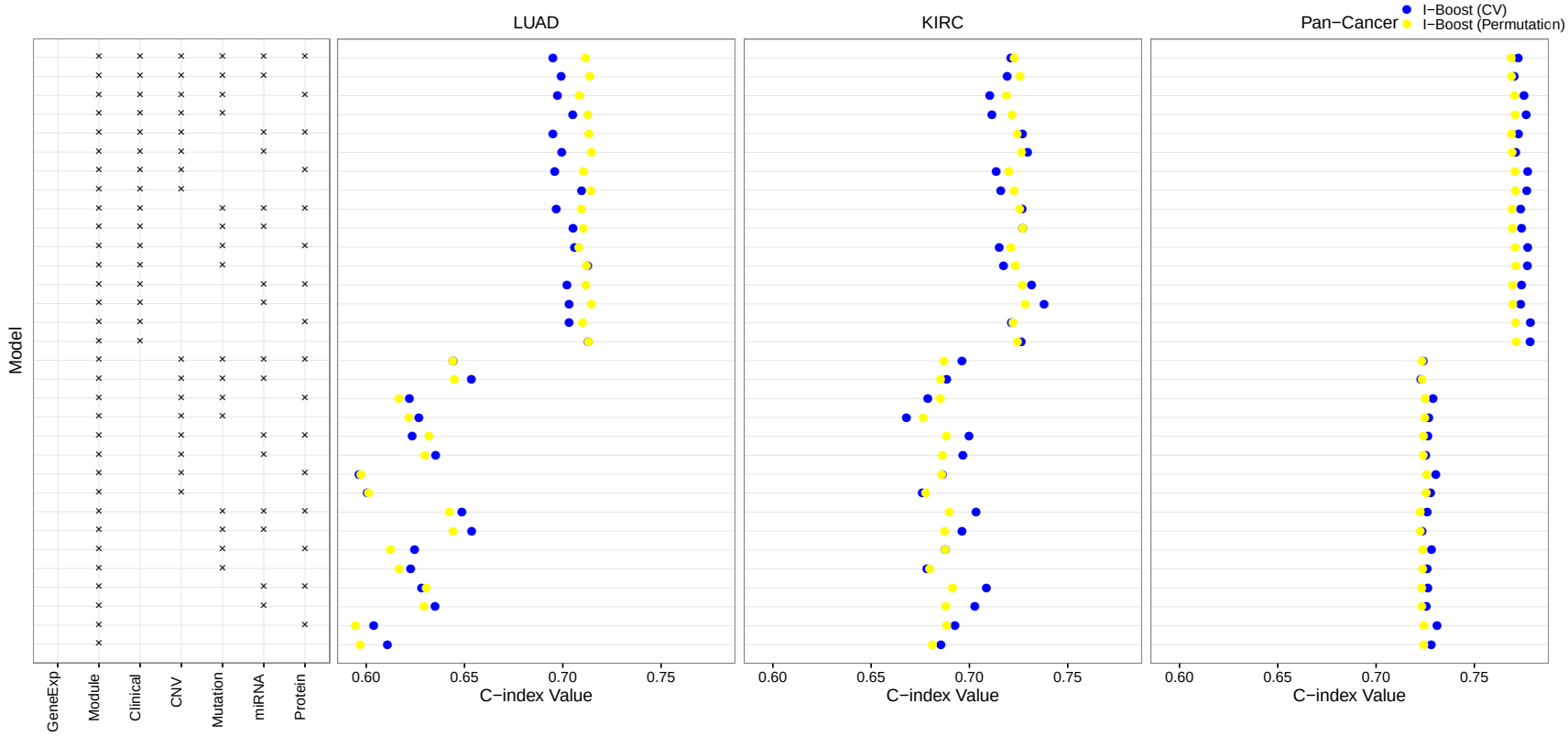


Figure 4.12. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using I-Boost-CV and I-Boost-Permutation for Models With Gene Modules. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing I-Boost-CV or I-Boost-Permutation on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

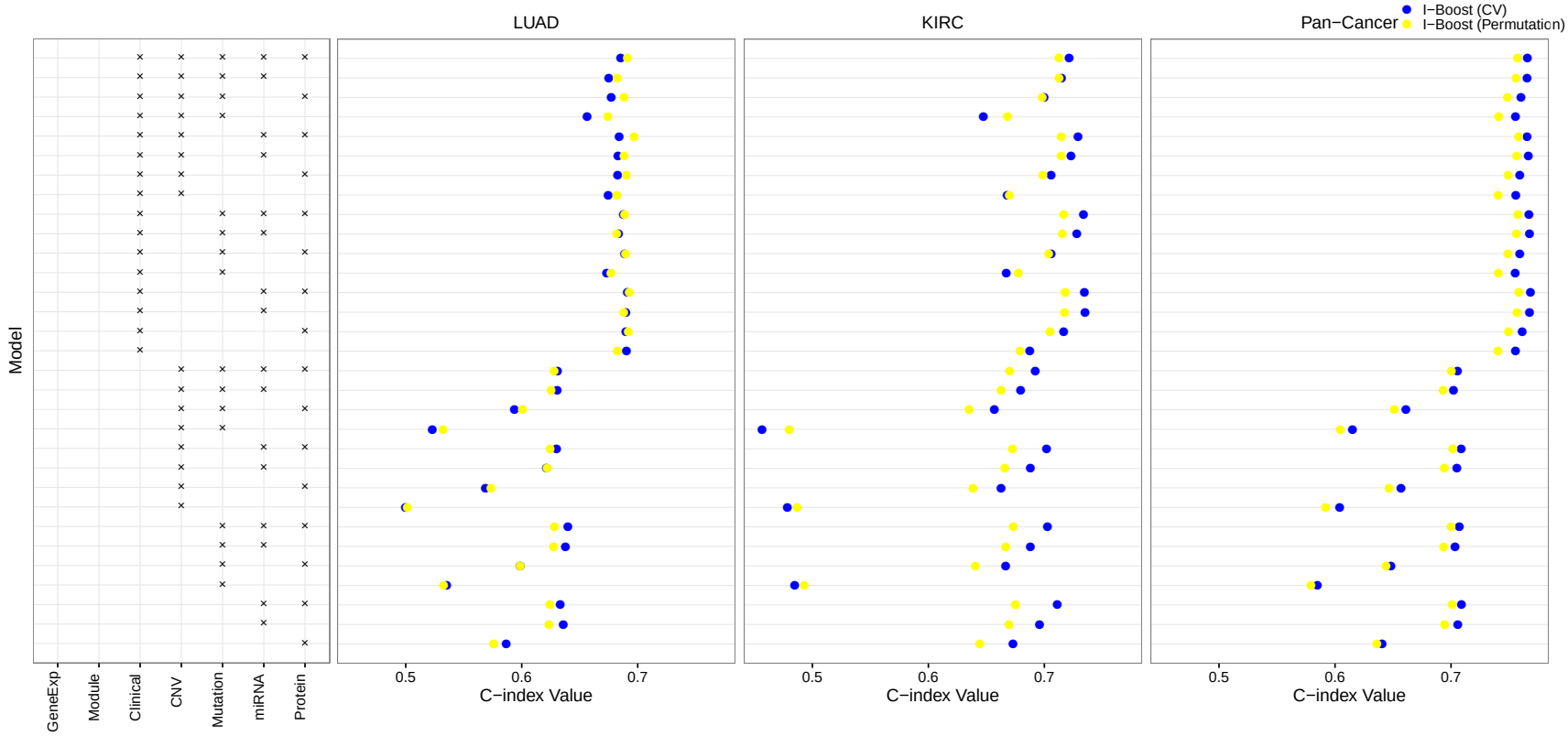


Figure 4.13. Analysis Results for the TCGA LUAD, KIRC, and Pan-Cancer Datasets, Using I-Boost-CV and I-Boost-Permutation for Models Without Gene Expression Data. Each row represents a particular combination of data types used as predictors, as indicated by the box on the left. Each dot is an average C-index value obtained by performing I-Boost-CV or I-Boost-Permutation on 30 training and testing dataset pairs. See the caption of Figure 4.1 for the abbreviations of the data types.

Table 4.1. Analysis Results From I-Boost-Permutation for the TCGA LUAD Dataset.

Predictor	Estimate
Module_UNC_MPYMT_NEU_Cluster_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.2581
Mutation_HMCN1	−0.0502
Mutation_FAT3	−0.0360
Clinical_gender_female_0	−0.0071
miRNA_hsa-miR-181c-5p	−0.0067
Mutation_AHNAK2	−0.0027
Mutation_LAMA2	−0.0002
CNV_BeroukhimS5.chr6:129217882-131730157.amp	0.0000
CNV_BeroukhimS2.19p12-66	0.0221
Module_UNC_Glycolysis_Signature_Median_BMC.Med.2009_PMID.19291283	0.0359
Module_IMMUNE_Bindea_Cell_Th2 cells_Median_Immunity.2013_PMID.24138885	0.0526
miRNA_hsa-miR-582-3p	0.1424
Clinical_age	0.1570
Clinical_pathologic_N	0.4687

NOTE: "Estimate" is the estimate of the log hazard ratio under the Cox proportional hazards model, where a positive value represents an increase of the hazard. The predictors are standardized to have unit standard deviation. Gender is coded as female = 0 and male = 1; pathologic stage T is dichotomized into T1 (0) and T2-T4 (1); pathologic stage N is dichotomized into N0 (0) and N1-N3 (1).

Age and pathological nodal status were selected as strong negative prognostic factors in the analyses of the LUAD, KIRC, and pan-cancer data sets. Age has been reported to be associated with prognosis for many cancer types (Lieu et al. 2014; de la Rochefordière et al. 1993; Asmis et al. 2008). In the analysis of the pan-cancer dataset, indicators of the multiple cancer types included in this dataset were selected, which is logical, since the survival time is known to depend on cancer types (Hoadley et al. 2014); the tissue of origin is an important prognostic factor. Among the gene modules, Glycolysis_signature and MUnknown_24 were selected as negative prognostic factors in the LUAD and pan-cancer datasets; these two modules are correlated with Hypoxia signatures (Pearson correlation = 0.59) among a set of 1,198 TCGA breast cancer samples. Likewise, Pcorr_IGS_Correlation and Activate_Endothelium, which were selected as negative prognostic factors for the pan-cancer dataset, were correlated with proliferation signatures (Pearson correlation = 0.96); these features are robust negative prognostic factors.

In contrast, signatures of CD8 T cells, non-inflammatory breast cancer (nIBC and MM_Red2,

Table 4.2. Analysis Results From I-Boost-Permutation for the TCGA KIRC Dataset.

Predictor	Estimate
Protein_AR	−0.1071
Module_IMMUNE_Bindea_Cell_CD8 T cells_Median_Immunity.2013_PMID.24138885	−0.0695
Module_Mature_LuminalUp_Median_Nat.Med.2009_PMID.19648928	−0.0675
Module_UNC_MM_Red2_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.0662
Module_GP7_Estrogen signaling: r=0.97	−0.0583
miRNA_hsa-miR-10b-3p	−0.0383
Module_UNC_HS_Green1_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.0380
miRNA_hsa-miR-192-5p	−0.0280
Protein_Src_pY416	−0.0278
miRNA_hsa-miR-425-3p	−0.0167
Module_UNC_LUMINAL_Cluster_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.0160
Module_UNC_HS_Green8_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.0115
Protein_PRAS40_pT246	−0.0108
Module_UNC_Duke_Module06_er_Median_Mike_PMID:20335537	−0.0027
Module_Pcorr_squamoid_PLOS.2012_PMID.22590557	0.0049
Clinical_pathologic_N	0.0062
miRNA_hsa-miR-21-5p	0.0071
Module_UNC_MM_p53null.Basal_Median_Genome.Biol.2013_PMID.24220145	0.0085
miRNA_hsa-miR-21-3p	0.0118
Protein_Caveolin-1	0.0144
Protein_TIGAR	0.0256
miRNA_hsa-miR-92b-3p	0.0277
miRNA_hsa-miR-223-3p	0.0325
miRNA_hsa-miR-130a-3p	0.0576
miRNA_hsa-miR-222-3p	0.0602
Protein_IGFBP2	0.0639
miRNA_hsa-let-7a-3p	0.0718
Clinical_age	0.1106
Module_UNC_Scorr_Basal_Correlation_JCO.2009_PMID.19204204	0.1346
Clinical_pathologic_T	0.2490

NOTE: See NOTE to Table 4.1.

Pearson correlation = 0.84), and luminal features (Mature_LuminalUp, HS_Green1, HS_Green8, LUMINAL_Cluster, Duke_Module06_er, Pcorr_Dasatinib_L_Correlation, GP7_estrogen signaling, and HS_Green18, Pearson correlation = 0.74) were selected as positive prognostic factors for the KIRC or pan-cancer datasets. The NEU_cluster module was selected as a strong positive predictor for the LUAD dataset, which is biologically significant because this module represents epithelial luminal cell differentiation and thus tracks more differentiated and lower grade lung cancers. These selected features, together with their biological implications, demonstrate the robustness of the I-Boost methodology.

Table 4.3. Analysis Results From I-Boost-Permutation for the TCGA Pan-Cancer Dataset.

Predictor	Estimate
Module_Pcorr_Dasatinib_L_Correlation_Cancer.Res.2007_PMID.17332353	−0.1515
Module_UNC_MS_CD44_DOWN_Median_PNAS.2009_PMID.19666588	−0.0450
Module_UNC_HS_Green18_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.0382
Module_UNC_MPYMT_NEU_Cluster_Median_ BMC.Med.Genomics.2011_PMID.21214954	−0.0350
Module_UNC_MNOtch4_Median_BMC.Med.Genomics.2011_PMID.21214954	−0.0293
miRNA_hsa-miR-101-3p	−0.0285
Module_IMMUNE_Bindea_Cell_CD8 T cells_Median_Immunity.2013_PMID.24138885	−0.0222
Module_Shipitsin_CD44_B_Median_Cancer.Cell.2007_PMID.17349583	−0.0184
Protein_p38_pT180_Y182	−0.0174
CNV_wa.9.p	−0.0146
Module_Inflammatory_Breast_Cancer_491_nIBC_CCR.2013_PMID.23396049	−0.0037
miRNA_hsa-miR-342-5p	−0.0016
miRNA_hsa-let-7b-5p	−0.0003
miRNA_hsa-miR-148a-3p	−0.0003
Protein_Dvl3	0.0001
Module_UNC_Glycolysis_Signature_Median_BMC.Med.2009_PMID.19291283	0.0007
Protein_PAI-1	0.0009
miRNA_hsa-miR-34a-5p	0.0010
miRNA_hsa-miR-511	0.0013

Continued on next page

Table 4.3 – continued from previous page

Predictor	Estimate
CNV_Basal.13q34-86	0.0081
Module_UNC_ADM_S100A10_A110NDGR1_Cluster_Median_ BMC.Med.Genomics.2011_PMID.21214954	0.0140
Module_Extensive_Residual_Diesase_ER54_Median_JAMA.2011_PMID.21558518	0.0172
Module_Pcorr_IGS_Correlation_NJEM.2007_PMID.17229949	0.0175
Clinical_gender_female_0	0.0292
Module_UNC_Activate.Endothelium_Median_Clin.Exp.Metastasis.2014_PMID.23975155	0.0311
Module_UNC_MUnknown_24_Median_BMC.Med.Genomics.2011_PMID.21214954	0.0447
miRNA_hsa-miR-582-3p	0.0571
Clinical_LUAD	0.0758
Clinical_HNSC	0.0762
Clinical_BLCA	0.0805
Clinical_KIRC	0.0916
Module_UNC_Duke_Module20_stat3_Median_Mike_PMID:20335537	0.1102
Clinical_pathologic_N	0.1739
Clinical_pathologic_T	0.1949
Clinical_age	0.3289

NOTE: See NOTE to Table 4.1. For cancer type, BRCA is the reference group.

4.5 Discussion

In this chapter, we propose a novel method, termed I-Boost, for variable selection and outcome prediction that is especially powerful when one wishes to consider multiple data types at once. We used simulation studies and real-data analyses to demonstrate that in the presence of small but predictive data types (such as clinical variables), I-Boost produces better outcome prediction than LASSO and elastic net. In addition, I-Boost selects fewer variables than elastic net, which may be preferable for follow-up experiments. Finally, I-Boost is less sensitive to the tuning parameter selection procedure than elastic net.

Consistent with the current literature, we found that clinical variables are strong predictors of survival time. With I-Boost, we were able to build upon the clinical variables and extract additional useful information from genomic variables in order to improve the prediction; the improvement that

we obtained with I-Boost was considerably larger than that obtained by either LASSO or elastic net. We also compared the use of individual gene expression data versus gene modules and found that the use of gene modules leads to similar or better prediction accuracy and more interpretable results. When we considered the selected I-Boost models, clinical variables (e.g., age, tumor size, and pathological nodal status) were strong predictors of survival. I-Boost also selected several gene modules that were previously identified as prognostic of outcomes.

Our study has some limitations. The main limitation is that the LUAD and KIRC datasets pertain to a relatively small number of patients, with an even smaller number of observed deaths. This limitation motivated us to combine eight solid epithelial tumor types to form a large pan-cancer dataset. The analyses on the pan-cancer data might not properly account for heterogeneity across different cancer types. Another limitation of our study is that the quality of the clinical data differs across different cancer types; for example, the follow-up time for some cancer types was quite short.

For all analyses performed herein, the outcome of interest was the overall survival time. It may be preferable to use time-to-tumor progression as the outcome as it is more directly related to the clinical and genomic predictors. In that case, I-Boost needs to be extended to accommodate recurrent events, as a patient may experience multiple tumor progression events. Finally, the I-Boost methodology is applicable to any disease states where multiple types of genomic data are available.

4.6 Detailed Data Description

Data on 2,272 TCGA samples representing the eight different cancer types listed in Chapter 4.1 were obtained from the December 22, 2012 Pan-Cancer-12 data freeze from the Sage Bionetworks repository Synapse (https://www.synapse.org). We used the dataset that was previously processed and described by Hoadley et al. (2014) for all data types except protein expression. The protein data were downloaded from Broad GDAC Firehose (http://gdac.broadinstitute.org/) on June 26, 2017. In the analyses, COAD and READ were combined as one cancer type.

Clinical variables included gender, age, pathological stages T and N, and cancer type. For mRNA expression data, we used RNA-seq by Expectation-Maximization (RSEM) (Li and Dewey 2011) to quantify the transcript abundances measured by RNA sequencing and used the log2-transformed up-quantile-normalized RSEM values of 12,434 genes. The RNA sequencing was performed at the University of North Carolina at Chapel Hill (The Cancer Genome Atlas Research Network 2012*a*;*b*;*c*). Gene level expression data are also available on the TCGA Data Portal (https://tcga-

data.nci.nih.gov/tcga/). For mutation data, we used the single nucleotide variant calls, which were de-duplicated and re-annotated using the ENSEMBLE version 69 transcript database. A total of 130 genes with non-synonymous mutations in more than 10% of the whole sample were included for the analyses. The combined mutation annotation format file is available from the Synapse resource. For miRNA expression data, we used the read count data for 305 normalized expressions, which were compiled into an abundance matrix for 5p and 3p mature miRBase strands, as described by The Cancer Genome Atlas Research Network (2012*c*). For reverse-phase protein arrays, we used the level-3 normalized data for 136 proteins or phospho-proteins.. For copy number data, SNP6.0 array-based gene-level somatic copy number alteration data were generated from the GISTIC analysis (Zack et al. 2013). The input data matrix is available in Synapse at syn1710678. We used the copy number values for 216 cancer-specific segments, which are frequently aberrated in cancer of various types including breast cancer, and segments for all chromosome arms (a total of 41 segments) (Beroukhim et al. 2010; Chao et al. 2012).

We defined gene modules as sets of co-expressed genes that are considered to be functional units in breast cancer. We built a collection of 504 gene modules. The modules were constructed based on 73 publications or results from the Gene Set Enrichment Analysis (Subramanian et al. 2005). A partial list of the modules appears in Fan et al. (2011). Among the modules, 468 are median expression values for homogeneously expressed genes, 33 are correlations of expression values with predetermined gene centroids, and 3 are built from previously published gene expression prognostic models.

After removing patients with missing data, the total sample size was 1,420, including 202 LUAD samples and 195 KIRC samples. All survival times were censored at five years if the patients were still in the study at that time point. For the pan-cancer dataset, the median follow-up time was 16.8 months, and the censoring proportion was 77.6%. For the LUAD dataset, the median follow-up time was 13.9 months, and the censoring proportion was 71.3%. For the KIRC dataset, the median follow-up time was 28.9 months, and the censoring proportion was 63.6%.

CHAPTER 5

FUTURE WORK — VARIABLE SELECTION WITH MISSING DATA IN MULTI-PLATFORM GENOMICS STUDIES

5.1 Introduction

Most phenotypic variations of interest are not driven by the activities of a single gene but are the products of many genes acting in concert with one another at multiple layers of genomic structures. As a result, the marginal association of a certain feature of a gene with a phenotype may be quite different from the actual relationship between the gene and the phenotype. Also, the predictive power of a single gene is very low. It is therefore desirable to consider multiple types of genomic variables of all genes in a single analysis framework.

A naïve approach to model a phenotype and multiple types of genomic variables is to regress the phenotype on all genomic variables and model the relationships among each type of genomic variable using standard regression techniques. However, this approach is infeasible because each type of genomic variables is high-dimensional, and the number of parameters in the resulting model is much larger than the sample size of any genomic studies. The problem is further complicated by the presence of missing data. In a genomics study, due to cost constraints or other practical reasons, not all study subjects are measured for all genomic variables. For example, in The Cancer Genome Atlas (TCGA), protein expressions are only collected on a subset of subjects. Simple methods for handling missing data, such as complete-case analysis and single imputation, inevitably suffer efficiency loss and may even yield biased results. By contrast, the full-likelihood approach is valid, but it involves integration over all missing variables and thus is computationally infeasible for high-dimensional data.

In this chapter, we consider a high-dimensional regression model of a phenotype on multiple types of partially missing genomic variables and a latent variable model for the genomic variables. Under the proposed model, the likelihood function involves an integration of dimension that does not depend on the dimension of the genomic variables but depends only on the number of latent variables, which is chosen to be small. We develop a computationally efficient penalization EM

algorithm for variable selection.

In Chapter 5.2, we formulate the problem and develop the penalized estimator. In Chapter 5.3, we discuss the numerical implementation of the proposed estimator. In Chapter 5.4, we present preliminary results on the theoretical properties of the proposed estimator and outline their proofs.

5.2 Methods

Consider a genomics study that involves a phenotype Y, a p-dimensional vector of binary genomic variables $\boldsymbol{G}$, and a q-dimensional vector of continuous genomic variables $\boldsymbol{S}$. Both $\boldsymbol{G}$ and $\boldsymbol{S}$ may include multiple types of genomic variables. We assume the model:

$$\begin{aligned} Y \mid (\boldsymbol{G}, \boldsymbol{S}) &\sim f(Y; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{G} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi}) \\ \mathrm{logit}\{P(G_j = 1 \mid \boldsymbol{U})\} &= \boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{U}^* \quad \text{for } j = 1, \ldots, p \\ \boldsymbol{S} &= \boldsymbol{\Psi}\boldsymbol{U}^* + \boldsymbol{\epsilon}, \end{aligned}$$

where f is a density function with nuisance parameter $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are regression parameters, $\boldsymbol{U}^* = (1, \boldsymbol{U}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{U}$ is an r-dimensional multivariate standard normal latent variable, $\boldsymbol{\theta}_j \in \mathbb{R}^{r+1}$ is a vector of regression parameters, $\boldsymbol{\Psi} \in \mathbb{R}^{q \times (r+1)}$ is a matrix of regression parameters, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_q)^{\mathrm{T}}$, and ϵ_k follows i.i.d. $\mathrm{N}(0, \gamma_k^2)$. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^{\mathrm{T}}$. The models of $\boldsymbol{G}$ and $\boldsymbol{S}$ have been considered by Shen et al. (2009) for classifying patients into subtypes using multi-platform genomic data.

We allow elements of $\boldsymbol{G}$ and $\boldsymbol{S}$ to be missing. Let Q_j and R_j, by values 0 and 1, denote whether G_j and S_j are observed, respectively, $\boldsymbol{Q} = (Q_1, \ldots, Q_p)^{\mathrm{T}}$, and $\boldsymbol{R} = (R_1, \ldots, R_q)^{\mathrm{T}}$. The observed data consist of $(Y_i, \boldsymbol{Q}_i \circ \boldsymbol{G}_i, \boldsymbol{R}_i \circ \boldsymbol{S}_i, \boldsymbol{Q}_i, \boldsymbol{R}_i)$ for $i = 1, \ldots, n$, where $\circ$ denotes element-wise multiplication. Let $\boldsymbol{\nu} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\nu}) = \prod_{i=1}^{n} \sum_{\boldsymbol{G} \in \{0,1\}^p} \int\int & f(Y_i; \boldsymbol{\alpha}^{\mathrm{T}}\{\boldsymbol{Q}_i \circ \boldsymbol{G}_i + (1 - \boldsymbol{Q}_i) \circ \boldsymbol{G}\} \\ & + \boldsymbol{\beta}^{\mathrm{T}}\{\boldsymbol{R}_i \circ \boldsymbol{S}_i + (1 - \boldsymbol{R}_i) \circ (\boldsymbol{\Psi}\boldsymbol{U}^* + \boldsymbol{\epsilon})\}; \boldsymbol{\xi}) \\ & \times \prod_{j=1}^{p} \frac{e^{\{Q_{ij}G_{ij} + (1 - Q_{ij})G_j\}\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{U}^*}}{1 + e^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{U}^*}} \\ & \times \prod_{j=1}^{q} \frac{1}{\sqrt{\gamma_j^2}} e^{-\frac{1}{2\gamma_j^2}\{R_{ij}(S_{ij} - \boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{U}^*)^2 + (1 - R_{ij})\epsilon_j^2\}} e^{-\frac{1}{2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}} \, \mathrm{d}\boldsymbol{\epsilon} \, \mathrm{d}\boldsymbol{U}, \end{aligned}$$

where $\boldsymbol{G} = (G_1, \ldots, G_p)^{\mathrm{T}}$, and $\boldsymbol{\psi}_j^{\mathrm{T}}$ is the jth row of $\boldsymbol{\Psi}$.

To estimate the parameters, we adopt a penalized-likelihood approach with an (adaptive) L_1-penalty on $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and an L_2-penalty on $(\boldsymbol{\Theta}, \boldsymbol{\Psi})$. The penalized maximum likelihood estimator $\hat{\boldsymbol{\nu}} \equiv (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Psi}})$ is

$$\hat{\boldsymbol{\nu}} = \arg\max_{\boldsymbol{\nu}} \{\log L(\boldsymbol{\nu}) - \lambda_1(|\boldsymbol{w}_\alpha \circ \boldsymbol{\alpha}| + |\boldsymbol{w}_\beta \circ \boldsymbol{\beta}|) - \lambda_2(\|\boldsymbol{\Theta}^{(-1)}\|_2^2 + \|\boldsymbol{\Psi}^{(-1)}\|_2^2)\},$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^{\mathrm{T}}$, $(\boldsymbol{\Theta}^{(-1)}, \boldsymbol{\Psi}^{(-1)})$ is $(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ with the first columns removed, and $\boldsymbol{w}_\alpha$ and $\boldsymbol{w}_\beta$ are vectors of weights for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The weights $\boldsymbol{w}_\alpha$ and $\boldsymbol{w}_\beta$ consist of the inverse of the absolute values of $\tilde{\alpha}_j$ or $\tilde{\beta}_j$, where $\tilde{\alpha}_j$ and $\tilde{\beta}_j$ are some consistent estimators of α_j and β_j, respectively. In practice, we may set $\tilde{\alpha}_j$ and $\tilde{\beta}_j$ to be the L_1- or L_2-penalized maximum likelihood estimators or the maximum likelihood estimators of the regression coefficients in the models $f(Y; \alpha_j G_j; \boldsymbol{\xi})$ and $f(Y; \beta_j S_j; \boldsymbol{\xi})$, respectively.

5.3 Computation of the Penalized Estimator

We propose a penalization EM algorithm for the computation of $\hat{\boldsymbol{\nu}}$, with $\boldsymbol{U}$ and components of $\boldsymbol{\epsilon}$ that correspond to the missing components of $\boldsymbol{S}$ treated as missing data. For simplicity of presentation, assume that $\boldsymbol{G}$ is observed for all subjects and $\boldsymbol{S} = (\boldsymbol{S}^{(1)\mathrm{T}}, \boldsymbol{S}^{(2)\mathrm{T}})^{\mathrm{T}}$, such that $\boldsymbol{S}^{(1)}$ is observed for all subjects, and $\boldsymbol{S}^{(2)}$ is completely missing for some subjects. Let q_1 and q_2 be the dimensions of $\boldsymbol{S}^{(1)}$ and $\boldsymbol{S}^{(2)}$, respectively, and partition $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)\mathrm{T}}, \boldsymbol{\beta}^{(2)\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}^{(1)\mathrm{T}}, \boldsymbol{\epsilon}^{(2)\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\Psi} = (\boldsymbol{\Psi}^{(1)\mathrm{T}}, \boldsymbol{\Psi}^{(2)\mathrm{T}})^{\mathrm{T}}$ accordingly. Let $\boldsymbol{\Gamma}_1 = c^{-1}(\beta_{q_1+1}\gamma_{q_1+1}, \ldots, \beta_q\gamma_q)^{\mathrm{T}}$, $c = \{\sum_{j=q_1+1}^{q}(\beta_j\gamma_j)^2\}^{1/2}$, and $\boldsymbol{\Gamma}$ be an orthonormal matrix with the first row being $\boldsymbol{\Gamma}_1^{\mathrm{T}}$. Define $\tilde{\boldsymbol{\epsilon}}^{(2)} = \boldsymbol{\Gamma}\mathrm{diag}(\gamma_{q_1+1}^{-1}, \ldots, \gamma_q^{-1})\boldsymbol{\epsilon}^{(2)}$, such that $\tilde{\epsilon}_1^{(2)} = c^{-1}\boldsymbol{\beta}^{(2)\mathrm{T}}\boldsymbol{\epsilon}^{(2)}$. In terms of $\boldsymbol{U}$ and $\tilde{\boldsymbol{\epsilon}}^{(2)}$, the complete-data likelihood for the ith subject is

$$f(Y_i; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{G}_i + \boldsymbol{\beta}^{(1)\mathrm{T}}\boldsymbol{S}_i^{(1)} + \boldsymbol{\beta}^{(2)\mathrm{T}}\boldsymbol{\Psi}^{(2)}\boldsymbol{U}_i^* + c\tilde{\epsilon}_{i1}^{(2)}; \boldsymbol{\xi}) \prod_{j=1}^{p} \frac{e^{G_{ij}\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{U}_i^*}}{1 + e^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{U}_i^*}}$$
$$\times \prod_{j=1}^{q_1} \frac{1}{\sqrt{\gamma_j^2}} e^{-\frac{1}{2\gamma_j^2}(S_{ij} - \boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{U}_i^*)^2} e^{-\frac{1}{2}\tilde{\epsilon}_{i1}^{(2)2}} \prod_{j=2}^{q_2} e^{-\frac{1}{2}\tilde{\epsilon}_{ij}^{(2)2}} e^{-\frac{1}{2}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i}.$$

Therefore, $(\boldsymbol{U}, \tilde{\epsilon}_1^{(2)})$ is independent of $\tilde{\boldsymbol{\epsilon}}_{-1}^{(2)}$, and $\tilde{\boldsymbol{\epsilon}}_{-1}^{(2)}$ follows the $(q_2 - 1)$-dimensional multivariate standard normal distribution, where $\tilde{\boldsymbol{\epsilon}}_{-1}^{(2)}$ is a vector of the last $(q_2 - 1)$ elements of $\tilde{\boldsymbol{\epsilon}}^{(2)}$. The posterior expectation of any function of $\boldsymbol{U}$ can be calculated by numerical integration over $(\boldsymbol{U}, \tilde{\epsilon}_1^{(2)})$, instead

of the whole set of high-dimensional missing data.

Upon calculation of the weights $\boldsymbol{w}_\alpha$ and $\boldsymbol{w}_\beta$, the penalization EM algorithm iterates over the following steps until convergence:

1. For each subject, obtain the weights for the Gauss-Hermite quadrature for the conditional expectation of any function of $\boldsymbol{U}$. Calculate the first and second moments of $(\boldsymbol{U}, \tilde{\epsilon}_1^{(2)})$ conditional on the observed data, and use them to obtain the first and second conditional moments of $(\boldsymbol{U}, \boldsymbol{\epsilon}^{(2)})$. Let $\hat{\mathrm{E}}$ be the conditional expectation with respect to $(\boldsymbol{U}, \boldsymbol{\epsilon}^{(2)})$.

2. For $j = 1, \ldots, p$, maximize

$$\sum_{i=1}^{n} \hat{\mathrm{E}}\{\boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{U}_i^* G_{ij} - \log(1 + e^{\boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{U}_i^*})\} - \lambda_2 \|\boldsymbol{\theta}_j^{(-1)}\|_2^2$$

   using the Newton-Raphson algorithm, where $\boldsymbol{\theta}_j^{(-1)}$ is $\boldsymbol{\theta}_j$ with the first element removed.

3. For $j = 1, \ldots, q$, maximize

$$\sum_{i=1}^{n} -\frac{R_{ij}}{2} \hat{\mathrm{E}}(S_{ij} - \boldsymbol{\psi}_j^{\mathrm{T}} \boldsymbol{U}_i^*)^2 - \lambda_2 \gamma_j^2 \|\boldsymbol{\psi}_j^{(-1)}\|_2^2$$

   at the current estimate of γ_j^2, where $\boldsymbol{\psi}_j^{(-1)}$ is $\boldsymbol{\psi}_j$ with the first element removed. A closed-form solution is available for $\boldsymbol{\psi}_j$. Then, estimate γ_j^2 using the empirical sum of squares.

4. Let $\tilde{\boldsymbol{S}}_i = \boldsymbol{R}_i \circ \boldsymbol{S}_i + (1 - \boldsymbol{R}_i) \circ (\boldsymbol{\Psi} \boldsymbol{U}_i^* + \boldsymbol{\epsilon}_i)$, where $\boldsymbol{\Psi}$ is evaluated at the estimate from step (3). For the linear model, maximize

$$\sum_{i=1}^{n} Y_i \hat{\mathrm{E}}(\mu + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{G}_i + \boldsymbol{\beta}^{\mathrm{T}} \tilde{\boldsymbol{S}}_i) - \frac{1}{2} (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}) \begin{pmatrix} \boldsymbol{G}_i \boldsymbol{G}_i^{\mathrm{T}} & \boldsymbol{G}_i \hat{\mathrm{E}} \tilde{\boldsymbol{S}}_i^{\mathrm{T}} \\ \hat{\mathrm{E}} \tilde{\boldsymbol{S}}_i \boldsymbol{G}_i^{\mathrm{T}} & \hat{\mathrm{E}}(\tilde{\boldsymbol{S}}_i \tilde{\boldsymbol{S}}_i^{\mathrm{T}}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$$
$$-\lambda_1 \sigma^2 (|\boldsymbol{w}_\alpha \circ \boldsymbol{\alpha}| + |\boldsymbol{w}_\beta \circ \boldsymbol{\beta}|)$$

   using the "covariance-updates" algorithm of Friedman et al. (2010), where μ and σ^2 are the intercept and error variance, respectively. Then, update μ and σ^2 using the closed-form solutions.

5. For the generalized linear model or Cox proportional hazards model, replace the log-density

function of the phenotype in the log-likelihood function by a quadratic approximation (Friedman et al. 2010; Simon et al. 2011). The resulting objective function for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ takes the form

$$-\sum_{i=1}^{n} w_i \hat{\mathrm{E}}\{(z_i - \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{G}_i - \boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{S}}_i)^2\} - \lambda_1(|\boldsymbol{w}_\alpha \circ \boldsymbol{\alpha}| + |\boldsymbol{w}_\beta \circ \boldsymbol{\beta}|)$$

for some $(w_1, \ldots, w_n)$ and $(z_1, \ldots, z_n)$ that do not depend on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We can then update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using the covariance-updates method of step (4). Finally, update the nuisance parameters of the phenotype model using the Newton-Raphson algorithm or a closed-form solution, if available.

5.4 Preliminary Theoretical Results

In this section, we establish the theoretical properties of the proposed estimators under some rather restrictive settings. With an abuse of notation, let $\boldsymbol{\theta}$ denote the collection of all Euclidean parameters and $\boldsymbol{\eta}$ denote the collection of all infinite-dimensional parameters. Let $\mathbb{P}_n$ be the empirical measure and P be the true probability measure. Let $\ell(\boldsymbol{\theta}_n, \boldsymbol{\eta})$ be the log-likelihood function for one subject at $(\boldsymbol{\theta}_n, \boldsymbol{\eta})$, so that $n\mathbb{P}_n\ell(\boldsymbol{\theta}_n, \boldsymbol{\eta})$ is the log-likelihood function of a sample of size n, and $p\ell_n(\boldsymbol{\theta}_n) \equiv n \max_{\boldsymbol{\eta}} \mathbb{P}_n\ell(\boldsymbol{\theta}_n, \boldsymbol{\eta})$ is the profile log-likelihood. (We index $\boldsymbol{\theta}$ by the sample size n, such that the number of parameters may increase with the sample size.) To simplify notation, let $\boldsymbol{\beta}_n \in \mathbb{R}^{p_{1n}}$ denote the regression parameters of $(\boldsymbol{G}, \boldsymbol{S})$, $\boldsymbol{\psi}_n \in \mathbb{R}^{p_{2n}}$ denote the collection of regression parameters in the models of $\boldsymbol{G}$ and $\boldsymbol{S}$, $\boldsymbol{\xi}_n$ denote the remaining Euclidean parameters, and $\boldsymbol{\theta}_{n0} \equiv (\boldsymbol{\beta}_{n0}, \boldsymbol{\psi}_{n0}, \boldsymbol{\xi}_{n0}, \boldsymbol{\eta}_0)$ denote the true value of $(\boldsymbol{\beta}_n, \boldsymbol{\psi}_n, \boldsymbol{\xi}_n, \boldsymbol{\eta})$. Arrange $\boldsymbol{\beta}_n$ such that $\boldsymbol{\beta}_{n0} = (\boldsymbol{\beta}_{n0}^{(1)\mathrm{T}}, \boldsymbol{\beta}_{n0}^{(2)\mathrm{T}})^{\mathrm{T}}$, where the components of $\boldsymbol{\beta}_{n0}^{(1)}$ are non-zero, $\boldsymbol{\beta}_{n0}^{(2)} = \mathbf{0}$, and $\boldsymbol{\beta}_{n0}^{(1)}$ is of dimension s_n. Let p_n be the dimension of $\boldsymbol{\theta}_n$. The estimator $\hat{\boldsymbol{\theta}}_n$ maximizes the following penalized (profile) log-likelihood:

$$\Phi_n(\boldsymbol{\theta}_n) \equiv p\ell_n(\boldsymbol{\theta}_n) - n\lambda_{1n}\sum_{j=1}^{p_{1n}} w_{nj}|\beta_{nj}| - n\lambda_{2n}\sum_{j=1}^{p_{2n}} \psi_{nj}^2, \tag{5.1}$$

where $w_{nj} = |\tilde{\beta}_{nj}|^{-1}$, and $\tilde{\beta}_{nj}$ is an initial estimator of β_{nj}. Assume that:

(C1) The initial estimator $\tilde{\beta}_{nj} = O_p(n^{-\tau})$ for $s_n < j \le p_{1n}$ and some $\tau \le \frac{1}{2}$, and $|\tilde{\beta}_{nj}| > C_0 > 0$ $(1 \le j \le s_n)$ for large enough n and some constant C_0.

(C2) Each element of $\boldsymbol{\theta}_n$ belongs to some bounded subset of $\mathbb{R}$.

(C3) The tuning parameter $\lambda_{1n} = o(p_n^{-3/2} s_n^{-1/2})$ and $\lambda_{2n} = o(p_n^{-3/2} p_{2n}^{-1/2})$.

(C4) The number of parameters $p_n = o(n^{1/4})$.

Following Murphy and van der Vaart (2000), we assume the existence of an "approximately least-favorable submodel". For any fixed $(\boldsymbol{\theta}, \boldsymbol{\eta})$, let $\boldsymbol{t} \mapsto \boldsymbol{\eta}_{\boldsymbol{t}}(\boldsymbol{\theta}, \boldsymbol{\eta})$ be a map from the parameter space of $\boldsymbol{\theta}$ into the parameter space of $\boldsymbol{\eta}$. Let $l(\boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \ell(\boldsymbol{t}, \boldsymbol{\eta}_{\boldsymbol{t}}(\boldsymbol{\theta}, \boldsymbol{\eta}))$, such that l is three-times continuously differentiable almost surely with respect to the first component, and $\boldsymbol{l}^{(1)}$, $\boldsymbol{l}^{(2)}$, and $\boldsymbol{l}^{(3)}$ denote its the first, second, and third derivatives, respectively. The approximately least-favorable submodel satisfies the following conditions:

(C5) For every $(\boldsymbol{\theta}_n, \boldsymbol{\eta})$, $\boldsymbol{\eta}_{\boldsymbol{\theta}_n}(\boldsymbol{\theta}_n, \boldsymbol{\eta}) = \boldsymbol{\eta}$.

(C6) The first derivative $\boldsymbol{l}^{(1)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0) = \tilde{\boldsymbol{\ell}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)$, where $\tilde{\boldsymbol{\ell}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)$ is the efficient score statistic of $\boldsymbol{\theta}$.

(C7) For any random sequence $\tilde{\boldsymbol{\theta}}_n$ such that $\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n})$, $\hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n} \to_p \boldsymbol{\eta}_0$, where $\hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n} = \arg\max_{\boldsymbol{\eta}} \mathbb{P}_n \ell(\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\eta})$.

(C8) The efficient information matrix $\tilde{\boldsymbol{I}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0) \equiv P\tilde{\boldsymbol{\ell}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)\tilde{\boldsymbol{\ell}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)^{\mathrm{T}} = P\partial\tilde{\boldsymbol{\ell}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)/\partial\boldsymbol{\theta}_{n0}$, with $0 < C_1 < \lambda_{\min}\{\tilde{\boldsymbol{I}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)\} \leq \lambda_{\max}\{\tilde{\boldsymbol{I}}_n(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)\} < C_2 < \infty$ for all n and some constants C_1 and C_2, where $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalues of a matrix $\boldsymbol{A}$. The derivatives of l satisfy

$$
\begin{aligned}
&|P\boldsymbol{l}^{(1)}(\boldsymbol{\theta}_{n0}, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n})| \\
&\quad +|(\mathbb{P}_n - P)\{\boldsymbol{l}^{(1)}(\boldsymbol{\theta}_{n0}, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - \boldsymbol{l}^{(1)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)\}| &&= o_p(\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\| + n^{-1/2}) \\
&\qquad \|\mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_{n0}, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - \mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)\| &&= O_p(\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\|) \\
&\qquad\qquad P\{l_j^{(1)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0) l_k^{(1)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)\} &&< C_3 < \infty \\
&\qquad\qquad\qquad P\{l_{jk}^{(2)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_0)^2\} &&< C_4 < \infty,
\end{aligned}
$$

for some positive constants C_3 and C_4, and $|l_{jkh}^{(3)}(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}, \boldsymbol{\eta})|$ is bounded above by a random variable $\mathcal{M}$ for all values of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ in the parameter space, where $\mathrm{E}(\mathcal{M}^2) < C_5 < \infty$ for some positive constant C_5.

We have the following result.

Theorem 5.1. *Under conditions (C1)-(C8), there exists a local maximizer* $\hat{\boldsymbol{\theta}}_n$ *of* $\Phi_n(\boldsymbol{\theta}_n)$ *such that* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\| = O_p(\sqrt{p_n/n} + \sqrt{s}_n\lambda_{1n} + \sqrt{p_{2n}}\lambda_{2n})$.

Proof of Theorem 5.1. Let $\alpha_n = \sqrt{p_n/n} + \sqrt{s_n}\lambda_{1n} + \sqrt{p_{2n}}\lambda_{2n}$. It is sufficient to prove that for any $\epsilon > 0$, there exists a constant C such that

$$P\Big\{ \sup_{\|\boldsymbol{u}\|=C} \Phi_n(\boldsymbol{\theta}_{n0} + \alpha_n\boldsymbol{u}) < \Phi_n(\boldsymbol{\theta}_{n0}) \Big\} > 1 - \epsilon,$$

i.e., for any n, there exists a local maximum of Φ_n in the $C\alpha_n$-neighborhood of $\boldsymbol{\theta}_{n0}$ with probability $1 - \epsilon$. Let

$$\begin{aligned} D_n(\boldsymbol{u}) &= \Phi_n(\boldsymbol{\theta}_{n0} + \alpha_n\boldsymbol{u}) - \Phi_n(\boldsymbol{\theta}_{n0}) \\ &\le p\ell_n(\boldsymbol{\theta}_{n0} + \alpha_n\boldsymbol{u}) - p\ell_n(\boldsymbol{\theta}_{n0}) \\ &\quad -n\lambda_{1n}\sum_{j=1}^{s_n} w_j(|\beta_{n0j} + \alpha_n u_{\beta j}| - |\beta_{n0j}|) - n\lambda_{2n}\sum_{j=1}^{p_{2n}}\{(\psi_{n0j} + \alpha_n u_{\psi j})^2 - \psi_{n0j}^2\}, \quad (5.2) \end{aligned}$$

where $(u_{\beta 1}, \ldots, u_{\beta p_{1n}})$ and $(u_{\psi 1}, \ldots, u_{\psi p_{2n}})$ are elements of $\boldsymbol{u}$ corresponding to $\boldsymbol{\beta}_n$ and $\boldsymbol{\psi}_n$, respectively. Let $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_{n0} + \alpha_n\boldsymbol{u}$. The difference of the first two terms on the right-hand side of (5.2) is

$$\begin{aligned} p\ell_n(\tilde{\boldsymbol{\theta}}_n) - p\ell_n(\boldsymbol{\theta}_{n0}) &= n\mathbb{P}_n\ell(\tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - n\mathbb{P}_n\ell(\boldsymbol{\theta}_{n0}, \hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}_{n0}}) \\ &\le n\mathbb{P}_n\ell(\tilde{\boldsymbol{\theta}}_n, \boldsymbol{\eta}_{\tilde{\boldsymbol{\theta}}_n}(\tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - n\mathbb{P}_n\ell(\boldsymbol{\theta}_{n0}, \boldsymbol{\eta}_{\boldsymbol{\theta}_{n0}}(\tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n})) \\ &= n\mathbb{P}_n l(\tilde{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - n\mathbb{P}_n l(\boldsymbol{\theta}_{n0}, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}), \end{aligned}$$

where the inequality follows because, by condition (C5), the first term is unchanged, and the second term is reduced by replacing $\hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}_{n0}}$ by $\boldsymbol{\eta}_{\boldsymbol{\theta}_{n0}}(\tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n})$. By the Taylor's series expansion at the first argument of l, the above is equal to

$$\begin{aligned} &n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}}\mathbb{P}_n\boldsymbol{l}^{(1)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) + \frac{n}{2}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}}\mathbb{P}_n\boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n})(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}) \\ &\quad +\frac{n}{6}\{(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}}\mathbb{P}_n\boldsymbol{l}^{(3)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_n^*, \hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}_n^*})(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}}\}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}), \quad (5.3) \end{aligned}$$

where $\boldsymbol{\theta}_n^*$ is some value between $\boldsymbol{\theta}_{n0}$ and $\tilde{\boldsymbol{\theta}}_n$. The first term of (5.3) is equal to

$$\begin{aligned} & n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}} \mathbb{P}_n \boldsymbol{l}^{(1)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) \\ = \; & n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}} (\mathbb{P}_n - P) \boldsymbol{l}^{(1)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) + n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}} P \boldsymbol{l}^{(1)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) \\ = \; & n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0})^{\mathrm{T}} (\mathbb{P}_n - P) \tilde{\boldsymbol{\ell}}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) + \|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\| o_p(n^{1/2} + n\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\|) \\ = \; & O_p(\alpha_n \sqrt{n p_n}) \|\boldsymbol{u}\| + o_p(\alpha_n^2 n) \|\boldsymbol{u}\|. \end{aligned}$$

By condition (C8),

$$\|\mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - \mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\| = O_p(p_n^{3/2}) \alpha_n \|\boldsymbol{u}\|.$$

Also, for any $\varepsilon > 0$

$$\begin{aligned} & P(\|\mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0) + \tilde{\boldsymbol{I}}_0(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\| > \frac{\varepsilon}{p_n}) \\ \leq \; & \frac{p_n^2}{\varepsilon^2} \mathrm{E} \sum_{j,k=1}^{p_n} \{\mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0) - P \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\}^2 \\ = \; & O_p(\frac{p_n^4}{n}) = o_p(1), \end{aligned}$$

where the last line follows from condition (C8) and the dominated convergence theorem. Therefore, the second term of (5.3) is equal to

$$\begin{aligned} & -\frac{n}{2} \alpha_n^2 \boldsymbol{u}^{\mathrm{T}} \tilde{\boldsymbol{I}}_0 \boldsymbol{u} + \frac{1}{2} n \alpha_n^2 \boldsymbol{u}^{\mathrm{T}} \{\mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_n}) - \mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0) + \mathbb{P}_n \boldsymbol{l}^{(2)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\eta}_0) + \tilde{\boldsymbol{I}}_0\} \boldsymbol{u} \\ = \; & -\frac{n}{2} \alpha_n^2 \boldsymbol{u}^{\mathrm{T}} \tilde{\boldsymbol{I}}_0 \boldsymbol{u} + n \alpha_n^2 O_p(p_n^{3/2} \alpha_n) \|\boldsymbol{u}\|^3 + o_p(n \alpha_n^2) \|\boldsymbol{u}\|^2 \\ = \; & -\frac{n}{2} \alpha_n^2 \boldsymbol{u}^{\mathrm{T}} \tilde{\boldsymbol{I}}_0 \boldsymbol{u} + o_p(n \alpha_n^2) \|\boldsymbol{u}\|^2, \end{aligned}$$

where the last equality follows because $p_n^{3/2} \alpha_n = p_n^2 n^{-1/2} + p_n^{3/2} s_n^{1/2} \lambda_{1n} + p_n^{3/2} p_{2n}^{1/2} \lambda_{2n} \to 0$ by conditions (C3) and (C4). The third term of (5.3) is bounded above by

$$|\frac{1}{6} \sum_{j,k,h}^{p_n} \mathbb{P}_n l_{jkh}^{(3)} u_j u_k u_h| \alpha_n^3 \leq O_p(p_n^{3/2} \alpha_n) n \alpha_n^2 \|\boldsymbol{u}\|^2 = o_p(n \alpha_n^2) \|\boldsymbol{u}\|^3.$$

The first penalty term, i.e., the third term of the right-hand side of (5.2), is bounded above by

$$\begin{aligned} n\lambda_{1n}\sum_{j=1}^{s_n} w_j(|\beta_{n0j}+\alpha_n u_{\beta j}|-|\beta_{n0j}|) &\leq n\lambda_{1n}\alpha_n\sum_{j=1}^{s_n} w_j|u_{\beta j}| \\ &\leq n\lambda_{1n}\alpha_n\|\boldsymbol{u}\|(\sum_{j=1}^{s_n} w_j^2)^{1/2} \\ &= O_p(n\lambda_{1n}\alpha_n\sqrt{s_n})\|\boldsymbol{u}\| = O_p(n\alpha_n^2)\|\boldsymbol{u}\|. \end{aligned}$$

The forth term of (5.2) is bounded above by

$$O_p(n\lambda_{2n}\alpha_n\sqrt{p_{2n}})\|\boldsymbol{u}\| + n\lambda_{2n}\alpha_n^2\|\boldsymbol{u}\|^2 = O_p(n\alpha_n^2)\|\boldsymbol{u}\| + o_p(n\alpha_n^2)\|\boldsymbol{u}\|^2.$$

Combining the above results,

$$D_n(\boldsymbol{u}) = -\frac{1}{2}n\alpha_n^2\boldsymbol{u}^{\mathrm{T}}\check{\boldsymbol{I}}_0\boldsymbol{u} + O_p(n\alpha_n^2)\|\boldsymbol{u}\| + o_p(n\alpha_n^2)\|\boldsymbol{u}\|^2,$$

which is negative for large enough C. The desired result follows. □

By Theorem 5.1, if we choose $\lambda_{1n} = O\{\sqrt{p_n/(s_n n)}\}$ and $\lambda_{2n} = O\{\sqrt{p_n/(p_{2n} n)}\}$, then condition (C3) is satisfied, and $\hat{\boldsymbol{\theta}}_n$ converges at rate $\sqrt{n/p_n}$. Variable-selection consistency is given by the following result.

Theorem 5.2. *If $n^{-\tau}\min_{j\leq s_n}\beta_{n0j}/\lambda_{1n}\to\infty$, $\lambda_{1n}n^{\frac{1}{2}+\tau}/\sqrt{p_n}\to\infty$, and the conditions of Theorem 5.1 hold, then with probability tending to 1, there exists a local maximum of $\Phi_n(\boldsymbol{\theta}_n)$, $\hat{\boldsymbol{\theta}}_n$, with $\hat{\boldsymbol{\beta}}_n$ such that $\hat{\beta}_{n0j}\neq 0$ for $j\leq s_n$, and $\hat{\beta}_{n0j}=0$ for $j>s_n$.*

Proof of Theorem 5.2. First, we prove that $\hat{\beta}_{n0j}=0$ for $j>s_n$. In light of Theorem 5.1, it suffices to prove that for any $\|\boldsymbol{\beta}_n^{(1)}-\boldsymbol{\beta}_{n0}^{(1)}\|+\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_{n0}\|+\|\boldsymbol{\xi}_n-\boldsymbol{\xi}_{n0}\| = O_p(\sqrt{p_n/n})$ and any constant C,

$$\Phi_n\{(\boldsymbol{\beta}_n^{(1)\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}},\boldsymbol{\psi}_n,\boldsymbol{\xi}_n\} = \max_{\|\boldsymbol{\beta}_n^{(2)}\|\leq C\sqrt{p_n/n}}\Phi_n\{(\boldsymbol{\beta}_n^{(1)\mathrm{T}},\boldsymbol{\beta}_n^{(2)\mathrm{T}})^{\mathrm{T}},\boldsymbol{\psi}_n,\boldsymbol{\xi}_n\},$$

or, equivalently, that $\partial\Phi_n(\boldsymbol{\theta}_n)/\partial\beta_{nj}^{(2)}$ and $\beta_{nj}^{(2)}$ have different signs on $[-C\sqrt{p_n/n},C\sqrt{p_n/n}]\backslash\{0\}$. By

Taylor's series expansion,

$$\frac{\partial \Phi_n(\boldsymbol{\theta}_n)}{\partial \beta_{nj}^{(2)}} = \frac{\partial p\ell_n(\boldsymbol{\theta}_{n0})}{\partial \beta_{nj}^{(2)}} + \sum_k \frac{\partial^2 p\ell_n(\boldsymbol{\theta}_{n0})}{\partial \theta_{nk} \partial \beta_{nj}^{(2)}} (\theta_{nk} - \theta_{n0k}) - n\lambda_{1n} w_{2j} \text{sgn}(\beta_{nj}^{(2)}).$$

Under regularity conditions, we can show that the first two terms on the right-hand side of the above equation is $O_p(\sqrt{np_n})$. Therefore, the above equation is

$$-n^{1+\tau}\lambda_{1n} n^{-\tau} w_{2j} \text{sgn}(\beta_{nj}^{(2)}) + O_p(\sqrt{np_n}) = -n^{1+\tau}\lambda_{1n}\{n^{-\tau} w_{2j} \text{sgn}(\beta_{nj}^{(2)}) + O_p\{\sqrt{p_n}/(n^{\frac{1}{2}+\tau}\lambda_{1n})\}.$$

Because $n^{-\tau} w_{2_j}$ is bounded away from zero, the sign of the right-hand side above is $\text{sgn}(\beta_{nj}^{(2)})$ for large enough n.

Note that

$$\begin{aligned}
\min_{j \le s_n} |\hat{\beta}_{nj}| &\ge \min_{j \le s_n} |\beta_{n0j}| - \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| \\
&= \lambda_{1n} n^{-\tau} \{ \frac{n^\tau \min_{j \le s_n} |\beta_{n0j}|}{\lambda_{1n}} - O_p(n^{-\tau - \frac{1}{2}} p_n^{1/2} \lambda_{1n}) \} \\
&= \lambda_{1n} n^{\tau} \{ \frac{\min_{j \le s_n} |\beta_{n0j}|}{n^\tau \lambda_{1n}} - o_p(1) \},
\end{aligned}$$

which is strictly positive for large enough n. Therefore, $\hat{\beta}_{n0j} \neq 0$ for $j \le s_n$. □

If we can find an n^τ-consistent estimator of $\boldsymbol{\beta}_n$ and choose $\lambda_{1n} = O\{\sqrt{p_n/(ns_n)}\}$ and $\lambda_{2n} = O\{\sqrt{p_n/(np_{2n})}\}$, then the adaptive LASSO estimator is consistent and selects all and only the relevant covariates asymptotically. It remains to find the estimator of $\boldsymbol{\beta}_n$. A possibility is to use ridge regression to obtain the initial estimator of $\boldsymbol{\beta}_n$, i.e., $\tilde{\boldsymbol{\beta}}_n$ maximizes

$$p\ell_n(\boldsymbol{\theta}_n) - n\tilde{\lambda}_{1n} \sum_{j=1}^{p_{1n}} \beta_{nj}^2 - n\tilde{\lambda}_{2n} \sum_{j=1}^{p_{2n}} \psi_{nj}^2.$$

By the same arguments as the proof of Theorem 5.1, $\|\tilde{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_{n0}\| = O_p(\sqrt{p_n/n})$, if $\tilde{\lambda}_{1n} = O\{\sqrt{p_n/(p_{1n}n)}\}$, and $\tilde{\lambda}_{2n} = O\{\sqrt{p_n/(p_{2n}n)}\}$.

BIBLIOGRAPHY

Ahmed, A. A., Mills, A. D., Ibrahim, A. E., Temple, J., Blenkiron, C., Vias, M., Massie, C. E., Iyer, N. G., McGeoch, A., Crawford, R. et al. (2007), "The Extracellular Matrix Protein TGFBI Induces Microtubule Stabilization and Sensitizes Ovarian Cancers to Paclitaxel," *Cancer Cell*, 12, 514–527.

Arend, R. C., Londoño-Joshi, A. I., Straughn, J. M., and Buchsbaum, D. J. (2013), "The Wnt/β-Catenin Pathway in Ovarian Cancer: A Review," *Gynecologic Oncology*, 131, 772–779.

Asmis, T. R., Ding, K., Seymour, L., Shepherd, F. A., Leighl, N. B., Winton, T. L., Whitehead, M., Spaans, J. N., Graham, B. C., and Goss, G. D. (2008), "Age and Comorbidity as Independent Prognostic Factors in the Treatment of Non–Small-Cell Lung Cancer: A Review of National Cancer Institute of Canada Clinical Trials Group Trials," *Journal of Clinical Oncology*, 26, 54–59.

Asparouhov, T., Masyn, K., and Muthen, B. (2006), "Continuous Time Survival in Latent Variable Models," in *Proceedings of the Joint Statistical Meeting in Seattle, August 2006. ASA Section on Biometrics*, pp. 180–187.

Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., Zhang, G., Franceschini, N., Fox, K., Lange, E. M. et al. (2012), "Imputation of Exome Sequence Variants Into Population-Based Samples and Blood-Cell-Trait-Associated Loci in African Americans: NHLBI GO Exome Sequencing Project," *The American Journal of Human Genetics*, 91, 794–808.

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G. et al. (2002), "Gene-Expression Profiles Predict Survival of Patients With Lung Adenocarcinoma," *Nature Medicine*, 8, 816–824.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M. et al. (2010), "The Landscape of Somatic Copy-Number Alteration Across Human Cancers," *Nature*, 463, 899–905.

Bollen, K. A. (1989), *Structural Equations With Latent Variables*, New York: Wiley.

Bollen, K. A., and Davis, W. R. (2009), "Two Rules of Identification for Structural Equation Models," *Structural Equation Modeling*, 16, 523–536.

Bühlmann, P. (2006), "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34, 559–583.

Bühlmann, P., and Yu, B. (2006), "Sparse Boosting," *Journal of Machine Learning Research*, 7, 1001–1024.

Caburet, S., Anttonen, M., Todeschini, A. L., Unkila-Kallio, L., Mestivier, D., Butzow, R., and Veitia, R. A. (2015), "Combined Comparative Genomic Hybridization and Transcriptomic Analyses of Ovarian Granulosa Cell Tumors Point to Novel Candidate Driver Genes," *BMC Cancer*, 15: 251.

Chao, H.-H., He, X., Parker, J. S., Zhao, W., and Perou, C. M. (2012), "Micro-Scale Genomic DNA Copy Number Aberrations as Another Means of Mutagenesis in Breast Cancer," *PLoS ONE*, 7: e51719.

Chareka, P. (2007), "A Finite-Interval Uniqueness Theorem for Bilateral Laplace Transforms,"

International Journal of Mathematics and Mathematical Sciences, 2007.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B*, 34, 187–220.

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C., Machiels, J.-P., Haustermans, K., and De Moor, B. (2009), "A Kernel-Based Integration of Genome-Wide Data for Clinical Decision Support," *Genome Medicine*, 1: 39.

Dahly, D., Adair, L., and Bollen, K. (2009), "A Structural Equation Model of the Developmental Origins of Blood Pressure," *International Journal of Epidemiology*, 38, 538–548.

De Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer.

de la Rochefordière, A., Campana, F., Fenton, J., Vilcoq, J., Fourquet, A., Asselain, B., Scholl, S., Pouillart, P., Durand, J.-C., and Magdelenat, H. (1993), "Age as Prognostic Factor in Premenopausal Breast Carcinoma," *The Lancet*, 341, 1039–1043.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society: Series B*, 39, 1–38.

Denis, M., and Tadesse, M. G. (2016), "Evaluation of Hierarchical Models for Integrative Genomic Analyses," *Bioinformatics*, 32, 738–746.

Denkert, C., Schmitt, W. D., Berger, S., Reles, A., Pest, S., Siegert, A., Lichtenegger, W., Dietel, M., and Hauptmann, S. (2002), "Expression of Mitogen-Activated Protein Kinase Phosphatase-1 (MKP-1) in Primary Human Ovarian Carcinoma," *International Journal of Cancer*, 102, 507–513.

Derkach, A., Lawless, J. F., and Sun, L. (2015), "Score Tests for Association Under Response-Dependent Sampling Designs for Expensive Covariates," *Biometrika*, 102, 988–994.

Evans, M., and Swartz, T. (2000), *Approximating Integrals Via Monte Carlo and Deterministic Methods*, New York: Oxford University Press.

Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., and Perou, C. M. (2011), "Building Prognostic Models for Breast Cancer Patients Using Clinical Variables and Hundreds of Gene Expression Signatures," *BMC Medical Genomics*, 4: 3.

Freund, Y., and Schapire, R. E. (1996), "Experiments With a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., and Beyene, J. (2009), "Data Integration in Genetics and Genomics: Methods and Challenges," *Human Genomics and Proteomics*, 2009: 869093.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Edition)*, New York: Springer.

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., and Li, H. (2013), "Sherlock:

Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS," *The American Journal of Human Genetics*, 92, 667–680.

Henderson, R., Diggle, P., and Dobson, A. (2000), "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics*, 1, 465–480.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V. et al. (2014), "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification Within and Across Tissues of Origin," *Cell*, 158, 929–944.

Hu, Y. J., Li, Y., Auer, P. L., and Lin, D. Y. (2015), "Integrative Analysis of Sequencing and Array Genotype Data for Discovering Disease Associations With Rare Mutations," *Proceedings of the National Academy of Sciences*, 112, 1019–1024.

Huang, J., and Wellner, J. A. (1997), "Interval Censored Survival Data: A Review of Recent Progress," in *Proceedings of the First Seattle Symposium in Biostatistics*, eds. D. Y. Lin, and T. R. Fleming, New York: Springer, pp. 123–169.

Huang, Y.-T. (2014), "Integrative Modeling of Multiple Genomic Data From Different Types of Genetic Association Studies," *Biostatistics*, 15, 587–602.

Huang, Y.-T. (2015), "Integrative Modeling of Multi-Platform Genomic Data Under the Framework of Mediation Analysis," *Statistics in Medicine*, 34, 162–178.

Huang, Y.-T., Cai, T., and Kim, E. (2016), "Integrative Genomic Testing of Cancer Survival Using Semiparametric Linear Transformation Models," *Statistics in Medicine*, 35, 2831–2844.

Huang, Y.-T., Liang, L., Moffatt, M. F., Cookson, W. O., and Lin, X. (2015), "iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis," *Genetic Epidemiology*, 39, 347–356.

Huang, Y.-T., and Pan, W.-C. (2015), "Hypothesis Test of Mediation Effect in Causal Mediation Model With High-Dimensional Continuous Mediators," *Biometrics*, 72, 402–413.

Huang, Y.-T., VanderWeele, T. J., and Lin, X. (2014), "Joint Analysis of SNP and Gene Expression Data in Genetic Association Studies of Complex Diseases," *The Annals of Applied Statistics*, 8, 352–376.

Jennings, E. M., Morris, J. S., Carroll, R. J., Manyam, G. C., and Baladandayuthapani, V. (2013), "Bayesian Methods for Expression-Based Integration of Various Types of Genomics Data," *EURASIP Journal on Bioinformatics and Systems Biology*, 2013: 13.

Killeen, A. P., Morris, D. G., Kenny, D. A., Mullen, M. P., Diskin, M. G., and Waters, S. M. (2014), "Global Gene Expression in Endometrium of High and Low Fertility Heifers During the Mid-Luteal Phase of the Estrous Cycle," *BMC Genomics*, 15: 234.

Kim, H., Golub, G. H., and Park, H. (2005), "Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation," *Bioinformatics*, 21, 187–198.

Kortram, R. A., van Rooij, A. C. M., Lenstra, A. J., and Ridder, G. (1995), "Constructive Identification of the Mixed Proportional Hazards Model," *Statistica Neerlandica*, 49, 269–281.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004), "Robust Inference for Univariate Proportional Hazards Frailty Regression Models," *The Annals of Statistics*, 32, 1448–1491.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014), "Principles and Methods of Integrative Genomic Analyses in Cancer," *Nature Reviews Cancer*, 14, 299–313.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004), "A Statistical Framework for Genomic Data Fusion," *Bioinformatics*, 20, 2626–2635.

Larsen, K. (2004), "Joint Analysis of Time-to-Event and Multiple Binary Indicators of Latent Classes," *Biometrics*, 60, 85–92.

Larsen, K. (2005), "The Cox Proportional Hazards Model With a Continuous Latent Variable Measured by Multiple Binary Indicators," *Biometrics*, 61, 1049–1055.

Lawless, J. (2016), "Two-Phase Outcome-Dependent Studies for Failure Times and Testing for Effects of Expensive Covariates," *Lifetime Data Analysis* [online], DOI:10.1007/s10985-016-9386-8. Available at *https://link.springer.com/journal/10985*.

Lee, S., Jhun, M., Lee, E.-K., and Park, T. (2007), "Application of Structural Equation Models to Construct Genetic Networks Using Differentially Expressed Genes and Single-Nucleotide Polymorphisms," *BMC Proceedings*, 1(Suppl 1): S76.

Li, B., and Dewey, C. N. (2011), "RSEM: Accurate Transcript Quantification From RNA-Seq Data With or Without a Reference Genome," *BMC Bioinformatics*, 12: 323.

Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., and Churchill, G. A. (2006), "Structural Model Analysis of Multiple Quantitative Traits," *PLoS Genetics*, 2: e114.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010), "MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes," *Genetic Epidemiology*, 34, 816–834.

Lieu, C. H., Renfro, L. A., de Gramont, A., Meyers, J. P., Maughan, T. S., Seymour, M. T., Saltz, L., Goldberg, R. M., Sargent, D. J., Eckhardt, S. G. et al. (2014), "Association of Age With Survival in Patients With Metastatic Colorectal Cancer: Analysis From the ARCAD Clinical Trials Program," *Journal of Clinical Oncology*, 32, 2975–2982.

Lin, D. Y., Zeng, D., and Tang, Z. Z. (2013), "Quantitative Trait Analysis in Sequencing Studies Under Trait-Dependent Sampling," *Proceedings of the National Academy of Sciences*, 110, 12247–12252.

Little, R. J. (1992), "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237.

Liu, Q., and Pierce, D. A. (1994), "A Note on Gauss-Hermite Quadrature," *Biometrika*, 81, 624–629.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013), "Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types," *The Annals of Applied Statistics*, 7, 523–542.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society: Series B*, 44, 226–233.

Menendez, J. A., and Lupu, R. (2007), "Fatty Acid Synthase and the Lipogenic Phenotype in Cancer Pathogenesis," *Nature Reviews Cancer*, 7, 763–777.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013), "Pattern Discovery and Cancer Gene Identification in Integrated Cancer Genomic Data," *Proceedings of the National Academy of Sciences*, 110, 4245–4250.

Moustaki, I., and Steele, F. (2005), "Latent Variable Models for Mixed Categorical and Survival Responses, With an Application to Fertility Preferences and Family Planning in Bangladesh," *Statistical Modelling*, 5, 327–342.

Murphy, S. A. (1994), "Consistency in a Proportional Hazards Model Incorporating a Random Effect," *The Annals of Statistics*, 22, 712–731.

Murphy, S. A., and van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Muthén, B., and Masyn, K. (2005), "Discrete-Time Survival Mixture Analysis," *Journal of Educational and Behavioral Statistics*, 30, 27–58.

Muthén, L., and Muthén, B. (1998–2015), *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Naliboff, B. D., Kim, S. E., Bolus, R., Bernstein, C. N., Mayer, E. A., and Chang, L. (2012), "Gastrointestinal and Psychological Mediators of Health-Related Quality of Life in IBS and IBD: A Structural Equation Modeling Analysis," *The American Journal of Gastroenterology*, 107, 451–459.

Nock, N. L., Larkin, E. K., Morris, N. J., Li, Y., and Stein, C. M. (2007), "Modeling the Complex Gene× Environment Interplay in the Simulated Rheumatoid Arthritis GAW15 Data Using Latent Variable Structural Equation Modeling," *BMC Proceedings*, 1(Suppl 1): S118.

Nock, N. L., Wang, X., Thompson, C. L., Song, Y., Baechle, D., Raska, P., Stein, C. M., and Gray-McGuire, C. (2009), "Defining Genetic Determinants of the Metabolic Syndrome in the Framingham Heart Study Using Association and Structural Equation Modeling Methods," *BMC Proceedings*, 3(Suppl 7): S50.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. et al. (2009), "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," *Journal of Clinical Oncology*, 27, 1160–1167.

Parner, E. (1998), "Asymptotic Theory for the Correlated Gamma-Frailty Model," *The Annals of Statistics*, 26, 183–214.

Pencina, M. J., and D'Agostino, R. B. (2004), "Overall C as a Measure of Discrimination in Survival Analysis: Model Specific Population Value and Confidence Interval Estimation," *Statistics in Medicine*, 23, 2109–2123.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies,"

Biometrika, 66, 403–411.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004), "Generalized Multilevel Structural Equation Modeling," *Psychometrika*, 69, 167–190.

Rabe-Hesketh, S., Yang, S., and Pickles, A. (2001), "Multilevel Models for Censored and Latent Responses," *Statistical Methods in Medical Research*, 10, 409–427.

Reilly, T., and O'Brien, R. M. (1996), "Identification of Confirmatory Factor Analysis Models of Arbitrary Complexity: The Side-By-Side Rule," *Sociological Methods & Research*, 24, 473–491.

Sabourin, J. A., Valdar, W., and Nobel, A. B. (2015), "A Permutation Approach for Selecting the Penalty Parameter in Penalized Model Selection," *Biometrics*, 71, 1185–1194.

Schumaker, L. (2007), *Spline Functions: Basic Theory*, Cambridge: Cambridge University Press.

Seoane, J. A., Day, I. N., Gaunt, T. R., and Campbell, C. (2014), "A Pathway-Based Data Integration Framework for Prediction of Disease Progression," *Bioinformatics*, 30, 838–845.

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M.-S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Giordano, T. J., Misek, D. E. et al. (2008), "Gene Expression–Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study," *Nature Medicine*, 14, 822–827.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009), "Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis," *Bioinformatics*, 25, 2906–2912.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S. et al. (2002), "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning," *Nature Medicine*, 8, 68–74.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization Paths for Cox's Proportional Hazards Model Via Coordinate Descent," *Journal of Statistical Software*, 39, 1–13.

Stoolmiller, M., and Snyder, J. (2006), "Modeling Heterogeneity in Social Interaction Processes Using Multilevel Survival Analysis," *Psychological Methods*, 11, 164–177.

Stoolmiller, M., and Snyder, J. (2013), "Embedding Multilevel Survival Analysis of Dyadic Social Interaction in Structural Equation Models: Hazard Rates as Both Outcomes and Predictors," *Journal of Pediatric Psychology*, 39, 222–232.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., De Grassi, A., Lee, C. et al. (2007), "Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes," *Science*, 315, 848–853.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. et al. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, 102, 15545–15550.

Tang, Z. Z., and Lin, D. Y. (2013), "MASS: Meta-Analysis of Score Statistics for Sequencing Studies," *Bioinformatics*, 29, 1803–1805.

The Cancer Genome Atlas Research Network (2011), "Integrated Genomic Analyses of Ovarian Carcinoma," *Nature*, 474, 609–615.

The Cancer Genome Atlas Research Network (2012*a*), "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers," *Nature*, 489, 519–525.

The Cancer Genome Atlas Research Network (2012*b*), "Comprehensive Molecular Characterization of Human Colon and Rectal Cancer," *Nature*, 487, 330–337.

The Cancer Genome Atlas Research Network (2012*c*), "Comprehensive Molecular Portraits of Human Breast Tumours," *Nature*, 490, 61–70.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Tillé, Y., and Matei, A. (2016), "Sampling: Survey Sampling," R Package Version 2.8.

Torres-García, W., Zhang, W., Runger, G. C., Johnson, R. H., and Meldrum, D. R. (2009), "Integrative Analysis of Transcriptomic and Proteomic Data of Desulfovibrio Vulgaris: A Non-Linear Model to Predict Abundance of Undetected Proteins," *Bioinformatics*, 25, 1905–1914.

Tsiatis, A. A., and Davidian, M. (2004), "Joint Modeling of Longitudinal and Time-To-Event Data: An Overview," *Statistica Sinica*, 14, 809–834.

Tyekucheva, S., Marchionni, L., Karchin, R., and Parmigiani, G. (2011), "Integrating Diverse Genomic Data Using Gene Sets," *Genome Biology*, 12: R105.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002), "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature*, 415, 530–536.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010), "Inference of Patient-Specific Pathway Activities From Multi-Dimensional Cancer Genomics Data Using PARADIGM," *Bioinformatics*, 26, 237–245.

Vicard, P. (2000), "On the Identification of a Single-Factor Model With Correlated Residuals," *Biometrika*, 87, 199–205.

Wang, P. H., Lee, W. L., Juang, C. M., Yang, Y. H., Lo, W. H., Lai, C. R., Hsieh, S. L., and Yuan, C. C. (2005), "Altered mRNA Expressions of Sialyltransferases in Ovarian Cancers," *Gynecologic Oncology*, 99, 631–639.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013), "iBAG: Integrative Bayesian Analysis of High-Dimensional Multiplatform Genomics Data,"

Bioinformatics, 29, 149–159.

Wang, Z. Q., Bachvarova, M., Morin, C., Plante, M., Gregoire, J., Renaud, M. C., Sebastianelli, A., and Bachvarov, D. (2014), "Role of the Polypeptide N-Acetylgalactosaminyltransferase 3 in Ovarian Cancer Progression: Possible Implications in Abnormal Mucin O-Glycosylation," *Oncotarget*, 5, 544–560.

Ween, M. P., Oehler, M. K., and Ricciardelli, C. (2012), "Transforming Growth Factor-Beta-Induced Protein (TGFBI)/(βig-H3): A Matrix Protein With Dual Functions in Ovarian Cancer," *International Journal of Molecular Sciences*, 13, 10461–10477.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R., and Nevins, J. R. (2001), "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proceedings of the National Academy of Sciences*, 98, 11462–11467.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-Variant Association Testing for Sequencing Data With the Sequence Kernel Association Test," *The American Journal of Human Genetics*, 89, 82–93.

Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S., and Furey, T. S. (2012), "Integrating Genetic and Gene Expression Evidence Into Genome-Wide Association Analysis of Gene Sets," *Genome Research*, 22, 386–397.

Yu, B., Zheng, Y., Alexander, D., Morrison, A. C., Coresh, J., and Boerwinkle, E. (2014), "Genetic Determinants Influencing Human Serum Metabolome Among African Americans," *PLoS Genetics*, 10, e1004212.

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society: Series B*, 68, 49–67.

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L. A., Xu, Y., Hess, K. R., Diao, L. et al. (2014), "Assessing the Clinical Utility of Cancer Genomic and Proteomic Data Across Tumor Types," *Nature Biotechnology*, 32, 644–652.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H. et al. (2013), "Pan-Cancer Patterns of Somatic Copy Number Alteration," *Nature Genetics*, 45, 1134–1140.

Zeng, D., and Lin, D. Y. (2010), "A General Asymptotic Theory for Maximum Likelihood Estimation in Semiparametric Regression Models With Censored Data," *Statistica Sinica*, 20, 871–910.

Zhao, S. D., Cai, T. T., and Li, H. (2014), "More Powerful Genetic Association Testing Via a New Statistical Framework for Integrative Genomics," *Biometrics*, 70, 881–890.

Zhu, R., Zhao, Q., Zhao, H., and Ma, S. (2016), "Integrating Multidimensional Omics Data for Cancer Outcome," *Biostatistics*, 17, 605–618.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society: Series B*, 67, 301–320.